UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

WASHINGTON, DC 20460

August 25, 2000

OFFICE OF
ENVIRONMENTAL INFORMATION

MEMORANDUM

SUBJECT: Peer Review of *Guidance for Choosing a Sampling Design for Environmental Data Collection (EPA QA/G-5S)*

FROM: Nancy W. Wentworth /s/
Director, Quality Staff (2811R)

TO: Peer Review Panel

Attached is the peer review draft of the document *Guidance for Choosing a Sampling Design for Environmental Data Collection (EPA QA/G-5S)*. Your help in supplying peer review comments is greatly appreciated.

This document is intended to assist non-statistical analysts, managers, and decision makers in the selection of a suitable data collection strategy to meet planning objectives or documented criteria. Although statistical in nature, the guidance is written in plain English to ensure that the information can be found by those who can most effectively use it to improve data collection efforts.

The document contains two introductory/background chapters (1 and 2), with each subsequent chapter dedicated to a different sampling design. The chapters are written in a consistent format, such that each section answers a question that any analyst must address when selecting a sampling design. The appropriate use, strengths, and weaknesses of each design are discussed, as well as a description of the steps needed to implement each design. In order to make the chapters more readable by non-statisticians, the statistical theory and formulas for the implementation of the designs have been put in the appendix that follows each chapter. The overall goal of the guidance is to assure a positive answer to the question "Is this a representative sample from the population of interest?"

As a Peer Reviewer of this document, please given attention to some of the organizational aspects of this guidance (statistical Peer Reviewers will address the mathematical statistical aspects of the document). References to a particular section of a publication or handbook for relevant topics would be appreciated. Please comment on the following:

Overall:

- Does placing the more mathematical/statistical aspects of each sampling scheme in a self-contained appendix make the material more useable, or does it make the content too disjointed?

- Are the descriptions of the appropriate use of each design adequate for the analyst to determine the most appropriate design for the problem?

- Are the description of how to implement each design adequate and understandable?

In particular chapters:

- Do Chapters 1 and 2 provide sufficient background information so that the reader knows what this guidance can provide and how to find the information needed?

- Judgmental sampling (Chapter 3) is sometimes used for expediency with the data later (incorrectly) used as if the data came from a random sample. Please provide examples of any specific situations where this has been done but rectified at later stage; or provide information on problems that have resulted because of this type of mistake. Have you any further examples of benefits or limitations on the use of judgmental sampling?

- Chapters 4 and 5 concern Simple Random and Stratified Sampling respectively. Do these chapters contain sufficient information for a non-statistician to use these designs? Can these Chapters be expanded to meet potential user's needs?

- Chapter 6 concerns systematic and grid sampling, a practical method of obtaining environmental samples. Is there sufficient information to made this a practical aid to the non-statistician? Should additional algorithms by included for non-statistician or is the reference to Gilbert (1987) sufficient? Should systematic sampling as conducted at monitoring stations be included int his Chapter? Are there additional applications that should be addressed or used as examples in this chapter?

- Chapter 7, Ranked Set Sampling, is a relatively new topic that holds considerable promise as it marries together statistical rigor with objective input from the field scientist. Is this chapter clear on how to apply Ranked Set Sampling?

- Chapter 8, Adaptive Cluster Sampling, is another new topic that allows for directed samples to be merged with random samples when done correctly. Does the Chapter explain the concept adequately so that a non-statistician use it?

- The last Chapter, Composite Sampling, is by far the longest. This type of sampling

is commonly used when collecting a single sample from a small area, but less used on a large scale. Does this chapter explain the use of Composite Sampling sufficiently for non-statisticians? Is the illustration of the retesting schemes (Figure 28) clear?

Your comments and suggestions for improvement, together with any typographical errors or misstatements noted, should be emailed to John Warren at warren.john@epa.gov by October 30, 2000 or mailed to:

John Warren (2811R)
Quality Staff
U.S. Environmental Protection Agency
1200 Pennsylvania Ave, N.W.
Washington, D.C. 20460

Attachment

Guidance for Choosing a Sampling Design for Environmental Data Collection

Use in the Development of a
Quality Assurance Project Plan

EPA QA/G-5S

Quality Staff
Office of Environmental Information
United States Environmental Protection Agency

Washington, DC 20460

PEER REVIEW DRAFT

August 2000

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF BOXES

CHAPTER 1

INTRODUCTION

This document provides guidance on how to create sampling designs for environmental measurement data collection. A complete sampling design indicates the number and location of samples to be taken; in addition, it explains and justifies these decisions. This guidance describes several relevant basic and innovative sampling designs, and it describes the process for deciding which design is right for a particular application.

1.1 WHY IS SAMPLING DESIGN IMPORTANT?

The sampling design is a fundamental part of data collection for scientifically based decision making. A well-developed sampling design plays a critical role in ensuring that data are sufficient to draw the conclusions needed.[1]

A good, science-based decision requires high data quality. Measures of quality include the following: the appropriateness and accuracy of the sample collection and handling method, the quality and appropriateness of the laboratory analysis, and the representativeness of the data with respect to the object of the study. Representativeness is partly addressed through the sampling design. Consider Figure 1, a site map for a dry lagoon formerly fed by a pipe. Assuming that good field and laboratory practices are exercised and adequate quality control is implemented, the analytical results of soil samples drawn from randomly located sites A, B, and C may be representative if the object is to address whether the pipe has released a particular contaminant. However, these data are not representative if the object is to estimate the average concentration level of the entire old lagoon. For that estimation, random sampling locations should be generated from the entire site of the old lagoon (for example, perhaps including samples at D, E, and F). If the sampling design results in the collection of non-representative data, even the highest quality laboratory analysis cannot make up for the loss in data quality.

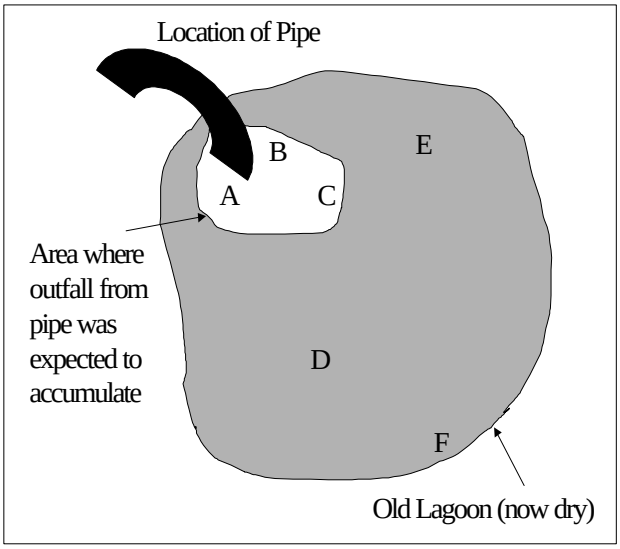


Figure 1. Site Map for Old Lagoon

[1]Note: The methods used in sample handling and extraction are equally important to data quality. United States Environmental Protection Agency (U.S. EPA) guidance on sampling methods and field sampling techniques are available to provide this information which will not be addressed in this document.

This guidance provides technical material on specific sampling designs that can be used to improve the quality of environmental data collected. Based in statistical theory, it explains the benefits and drawbacks of each design and describes relevant examples for illustration of environmental measurement applications. In order to choose a sampling design that adequately addresses the estimation or decision at hand, it is important to understand both how sampling design impacts data quality and what relevant factors should be considered. This guidance provides specific information on developing sampling designs that characterize the extent and level of contamination or help us decide whether a contaminant level exceeds a threshold or standard.

This guidance document provides the information needed to carry out the final step of the U.S. Environmental Protection Agency (U.S. EPA) Data Quality Objective Process (as described in *Guidance for the Data Quality Objectives Process (QA/G-4*) (U.S. EPA, 1994). This information is needed as a part of the Quality Assurance Project Plan (QA Project Plan) (as described in *Guidance for Quality Assurance Project Plans* (*QA/G-5)* (U.S. EPA, 1998b).

1.2 WHAT TYPES OF QUESTIONS WILL THIS GUIDANCE HELP TO ADDRESS?

Often it is difficult in practice to know how to answer many questions related to sampling design. This document will help to address the types of questions shown in Table 1.

Table 1. Questions that this Document Will Help to Address

• How much data do I need to make a good decision?
• How much data do I need to have an adequate estimate of the extent of contamination?
• How much data do I need to have a thorough search for a hot spot?
• How should I design my sampling to provide the best information for my problem given that I have a limited budget for sampling?
• Where should I take samples?
• Can the increased cost of more data collection be justified by the increase in the quality of the data?

1.3 WHO CAN BENEFIT FROM THIS DOCUMENT?

This document will be useful to anyone planning data collection from any type of environmental media, including soil, sediment, dust, surface water, groundwater, air, vegetation, and sampling in indoor environments. The potential benefits for different types of users are shown in Table 2.

Table 2. Potential Benefits for Users

Potential User	Benefit to the User
Environmental Scientist or **Environmental Engineer** who is planning the sampling or **Project Manager** planning the investigation and reviewing the sampling plan	An understanding of various sampling designs and the conditions under which these designs are appropriate. An understanding of how sampling design affects data quality and the ability to draw conclusions from the data. An understanding of the appropriate uses of professional judgment. The information needed to choose designs that may provide increased data quality at the same cost as compared to typical sampling approaches (e.g., Ranked Set Sampling).
Risk Assessor or **Data Analyst** who will be using the data	A full understanding of the advantages and limitations of data collected using various sampling designs. The ability to draw scientifically based conclusions from data based on different types of designs. The ability to match the assessment tools to the sampling design used.
Statistician assisting with the development and review of the sampling plan	Tables, figures, and text that will help communicate important information about choosing a sampling design to non-statistical colleagues working on the design. Advanced references to support more complex design development.

The information in this document is consistent with other U.S. EPA guidance documents on sampling design, including the Soil Screening Guidance (U.S. EPA, 1996) and SW-846 (U.S. EPA, 1986). This document is meant to apply to all environmental media; examples in this document will therefore be broader than in either of those specialized documents noted above. In addition, this document provides information on innovative designs not included in earlier documents.

Researchers must use systematic planning of data collection in order for them to draw scientifically based conclusions from collected data. For systematic planning of environmental data collection, U.S. EPA prefers the Data Quality Objectives (DQO) process described in the data quality objectives guidance document (U.S. EPA, 1994). A sampling design is chosen in Step 7 of the DQO process, based on the parameters specified in the other steps in the DQO process. This sampling design guidance explains the activities of Step 7 in Chapter 2, and a full discussion of the factors that should be considered in this step is given in Section 2.4.

1.4 HOW DOES THIS DOCUMENT FIT INTO THE U.S. EPA QUALITY SYSTEM?

Figure 2 illustrates the life-cycle of environmental data in the U.S. EPA Quality System. Existing guidance useful in navigating each step of the life-cycle of the data are identified by G-references. The process begins with systematic planning, the last step of which is sampling design. This is followed by the development of a QA Project Plan, which uses input from the sampling design and from data quality indicators.

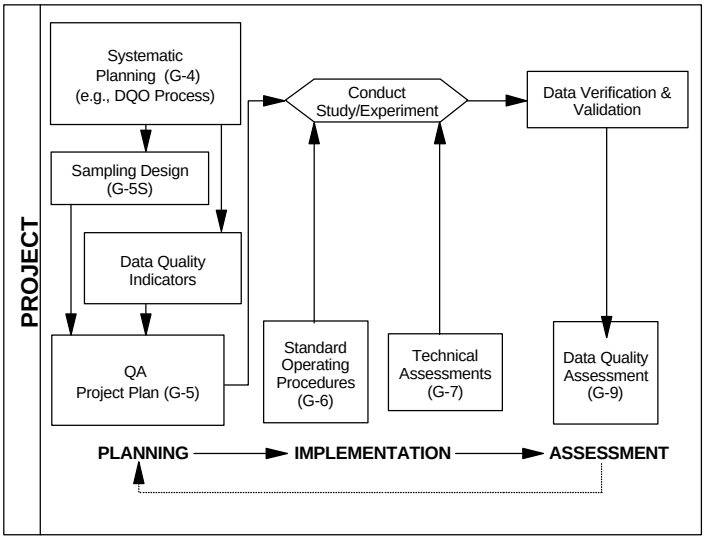


Figure 2. Life-cycle of data in the U.S. EPA Quality System

After the QA Project Plan is developed, data are collected during the Study/Experimental phase according to the plan. Quality is further assured by the use of standard operating procedures and audits (technical assessment). Finally, verification, validation, and quality assessment of the data complete the process.

1.5 HOW IS THIS DOCUMENT ORGANIZED?

After this chapter, most readers will want to continue on to Chapter 2. Chapter 2 defines important concepts and terms, provides an introduction to the types of sampling designs covered in this document, and discusses how input from systematic planning process affects the choice of a sampling design. Chapters 3 through 8 contain detailed information about six different sampling designs and sampling protocols. Some designs are often used in conjunction with other designs; descriptions and examples of these types of studies are included. In addition, there is a glossary defining key terms and a list of references containing citations for all referenced material and other materials used in the development of this document.

1.6 WHAT ARE THE LIMITATIONS OR CAVEATS TO THIS DOCUMENT?

The scope of this document is limited to environmental measurement data. Therefore, it does not explicitly address survey (questionnaire) data, human exposure data, or experimental data collection, although some of the concepts described here are applicable to these types of studies. This guidance does not provide a complete catalogue of the specific designs required by U.S. EPA for particular situations. These guidelines do not supercede regulatory requirements for specific types of sampling design nor regional, state, or program guidance; rather, they are intended to act in concert with other guidance.

CHAPTER 2

OVERVIEW OF SAMPLING DESIGN

Sampling design begins with a careful consideration of the objectives for collecting data. U.S. EPA advocates the use of the DQO Process to improve the quality of decisions made from environmental data. The DQO Process is a seven-step process to develop sampling designs for data collection activities that support decision making. The DQO Process is iterative and allows the planning team to incorporate new information and modify outputs from previous steps as inputs for a subsequent step. The DQO Process results in a design for collecting data (e.g., the number, type, and location of samples) and tolerable limits on the probabilities of making decision errors (false positives and false negatives).

By using the DQO Process, the planning team can clarify study objectives, define the appropriate type of data, and specify tolerable levels of potential decision errors that will be used to establish the quality and quantity of data needed to support decisions (U.S. EPA, 1994). Through this process, the planning team can examine trade-offs between uncertainty of results and cost of sampling and analysis in order to develop designs that are acceptable to all parties involved.

2.1 SAMPLING DESIGN CONCEPTS AND TERMS

Developing a sampling design is a crucial step in collecting appropriate and defensible data that accurately represent the problem being investigated. **Representativeness** is the correspondence between the analytical result and the actual environmental quality or condition experienced by a contaminant receptor (Barcelona, 1988, p. 12). Representativeness may be considered as the measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition (ANSI/ASQC E4-1994). Representative data are paramount to defensible decision making— if the data are not representative, then it matters little if they satisfy other Measurement Quality Objectives (goals or targets for accuracy, precision, completeness, or comparability of data).

A few additional key statistical terms are important to the understanding of sampling design. The **target population** is the set of units that comprise the items of interest in a scientific study, about which conclusions will be drawn. The **sampled population** is that part of the target population that is accessible and available for sampling. For example, the target population may be defined as surface soil in a residential yard, and the sampled population may be bare areas of soil in that yard not covered by structures or vegetation. Using a statistical design, inferences can be drawn about the target population, such as the concentration of fine particulates ($PM_{2.5}$) in ambient air in downtown Houston on a summer day, even though every single piece of the downtown air is not sampled. A **sampling unit** describes the members of the population that may be selected for sampling, such as individual trees or groups of trees. It is important for study

planners to take great care in defining a sampling unit's characteristics with respect to space and time. For example, a sampling unit may be defined as specific components of environmental media (e.g., cubic meters of air passing through a filter located in downtown Houston on July 15, 2000) or persons with particular characteristics (such as children between ages 3 and 12 living in Chicago during 1999).

Once a sampling unit is selected, a **measurement protocol** is applied; a measurement protocol is a specific procedure for making observations or performing analyses to determine the characteristics of interest for each sampling unit. The measurement protocol would include the procedures for collecting the physical sample, handling and preparing the physical sample, and applying an analytical method (including the sample preparation steps) to obtain a result (that is, to obtain the data from the sample). **Sample support** represents that portion of the sampling unit, such as an area, volume, or other quantity, that is extracted in the field and subjected to the measurement protocol. For example, if a sampling unit is a single tree, the sample support could be a core from the base of the tree. The **sampling design** specifies the number, type, and location (spatial or temporal) of sampling units to be selected for measurement.

A water sampling example illustrates how these terms relate to one another. Consider a study designed to measure *E. coli* and *entercocci* levels in a specific swimming area of a lake. The target population is the water flowing through this area (delineated by buoys) from May 1 until September 15. The sampled population will be the water in the swimming area at 7 a.m. and 2 p.m. The sampling unit chosen for the study is the water at 6 inches below the surface. The sample support is one liter of water. The measurement protocol calls for the use of a 2-liter beaker, held by a 6-inch handle. The sampler must use a non-motorized boat (e.g. a rowboat) to collect the sample so as to minimize the disturbance to the water. The sample is drawn in the specified and poured into a 2-liter sample jar, up to the 1-liter line. The rest of the water in the beaker is discarded back into the lake. Each 1-liter container of water is taken to the lab for analysis within 24 hours and is analyzed according to current state standards. The sampling design calls for taking a minimum of two samples on each sampling day at 7 a.m. and 2 p.m or up to three times a day when there are indications of increased potential for contamination (e.g., heavy rainfall). Sampling days are defined in the study and may be every day, every other day, or whichever level is appropriate for the particular problem at hand. The sampling design will also specify the exact location where the samples should be drawn.

As shown in Table 3, there are two main categories of sampling designs: probability-based designs and judgmental designs. **Probability-based sampling designs** apply sampling theory and involve random selection of sampling units. An essential feature of a probability sample is that each member of the population from which the sample was selected has a known probability of selection. When a probability-based design is used, **statistical inferences** may be made about the target population from the data obtained from the sampling units. **Judgmental sampling designs** involve the selection of sampling units on the basis of expert knowledge or professional judgment. See Chapter 3 on judgmental sampling for more details about the appropriate use of these designs.

Table 3. Probability-based versus Judgmental Sampling Designs

	Probability-based	Judgmental
Uses	• Estimating populations quantities • Detecting the presence of contamination with specific level of confidence	• Confirming presence or level of contamination at specific locations
Advantages	• Provides ability to calculate uncertainty associated with estimates • Provides reproducible results within uncertainty limits • Provides ability to make statistical inferences • Can handle decision error criteria	• Can be less expensive than probabilistic designs • Can be very efficient with knowledge of the site • Easy to implement
Disadvantages	• Can be more expensive than judgmental sampling • An optimal design depends on an accurate conceptual model	• Dependent upon expert knowledge • Cannot determine reliable estimates of variability • Interpretation of data relative to study objectives is dependent on personal judgment

Figure 3 illustrates the progression of the data collection process for both judgmental sampling and probabilistic sampling. Both processes start with defining the target population and the sampled population, and both end with the data collection and analysis. The difference is seen when moving up the diagram, in how conclusions can be drawn about the sampled and target populations.

With probabilistic sampling, the data analyst can use statistical tools to draw quantitative conclusions about the sampled population. That is, in estimating a parameter (e.g., the mean), the analyst can calculate the 95 percent confidence interval around the parameter of interest. In comparing to a threshold, the analyst can state whether the data indicate that the concentration exceeds or is below the threshold with a certain level of confidence. Professional judgment is then used to draw conclusions about the target population based on the statistical findings about the sampled population.

With judgmental sampling, statistical analysis cannot be used to draw conclusions about the target population, except on the basis of professional judgement. The usefulness of judgmental sampling will depend on the study objectives, the study size and scope, and the degree of professional judgment available. When judgmental sampling is used, quantitative statements about the level of confidence in an estimate (such as confidence intervals) cannot be made.

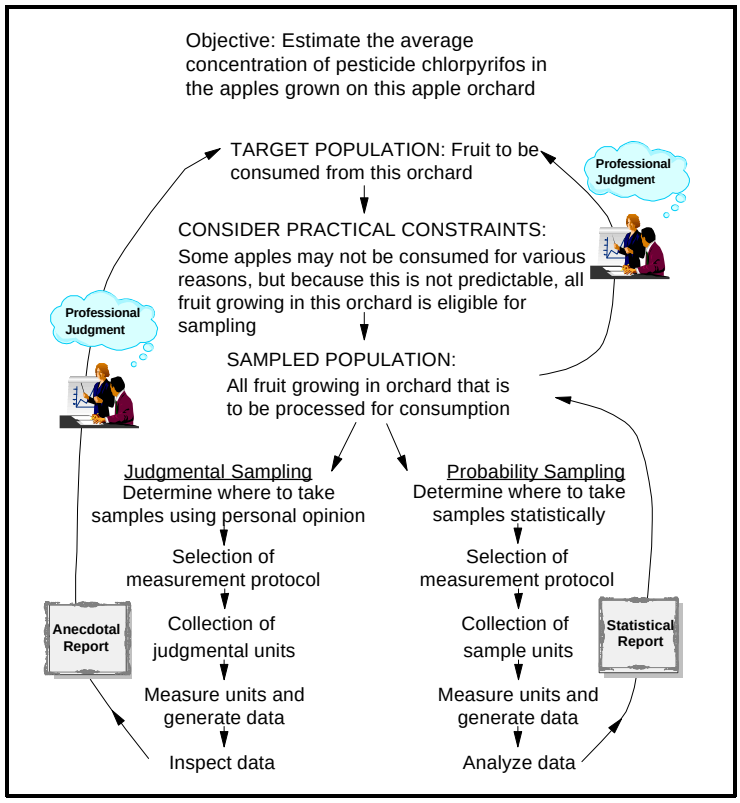


Figure 3. Inferences Drawn from Judgmental versus Probabilistic Sampling Designs

The final important concept for sampling design is the conceptual model. At the outset of data collection activities, it is critical to develop an accurate **conceptual model** of the potential hazard. A conceptual model describes the expected source of the contaminant and the size and breadth of the area of concern, identifies the relevant environmental media and the relevant fate and transport pathways, and defines the potential exposure pathways. It should also identify potential sources of variability in the data (e.g., inherent variability and sampling or analysis variance). Figure 4 shows an example of a conceptual model.

2.2 THE SAMPLING DESIGN PROCESS

What are the objectives of the sampling design process?

Sampling objectives can vary widely. The objective of a sampling design may be to support a decision about whether contamination levels exceed a threshold of unacceptable risk or whether certain characteristics of two populations differ by some amount, to estimate the mean characteristics of a population or the proportion of a population that has certain characteristics of interest, to identify the location of "hot spots" (areas having high levels of contamination), to characterize the nature and extent of contamination at a site, or to monitor trends in environmental conditions or indicators of health. A sampling design is intended to ensure that resulting data are adequately representative of the target population and defensible for their intended use. Throughout the sampling design process, a critical consideration is the efficient use of time, money, and human resources.

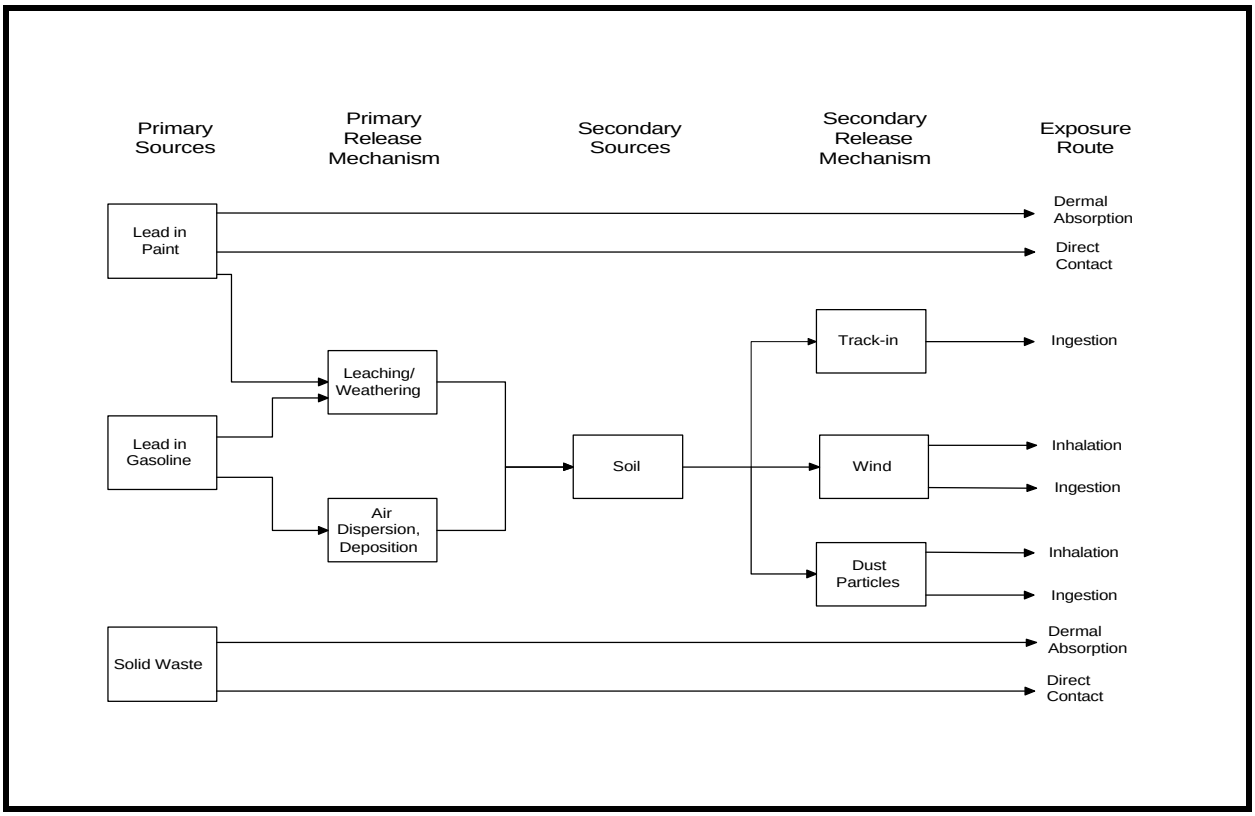


Figure 4. Example of a Conceptual Site Model of Risk from Lead in Soil

Who is typically involved in the sampling design process?

The sampling design process typically includes a multidisciplinary group that is involved in systematic planning (such as a DQO development team) at the beginning and at key review points. This team should include the decision maker or end user of the data. The more rigorous technical activities will likely be performed by statisticians or by environmental scientists or engineers who have training and experience in environmental statistics.

What information will be needed to implement the sampling design process?

Information needed includes outputs from the systematic planning process (for example, the outputs from Steps 1 through 6 of the DQO Process) and specific information about factors that should be considered in the choice of a sampling design, as described in the answer to the next question.

What factors affect the choice of a sampling design?

The categories of factors that should be used in the development of the design are outlined in Figure 5.

Information about the process or area of concern includes the conceptual model (described in Figure 4) and any additional information about the process or area (e.g., secondary data, pilot studies).

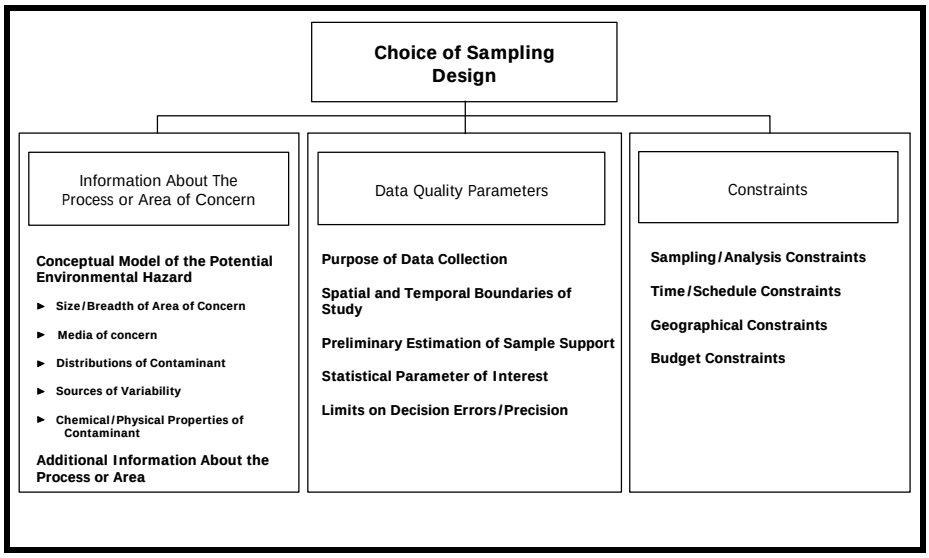


Figure 5. Criteria for Selecting a Sampling Design

Data Quality Parameters that are needed as input to the sampling design process are listed below. For more details on there parameters, the reader is referred to *U.S. EPA Guidance on the Data Quality Objectives Process* (U.S. EPA, 1994).

- The purpose of the data collection—that is, either **hypothesis testing** (evidence to reject or support a finding that a specific parameter exceeds a threshold level or evidence to reject or support a finding that the specified parameters of two populations are different), **estimating a parameter** with a level of confidence, or **detecting** hot spots (DQO Step 5).

- The spatial and temporal boundaries of the study (DQO Step 4).

- Preliminary estimation of sample support such as area, volume, or quantity that each sample represents (DQO Step 4).

- The statistical parameter of interest, such as mean, median, percentile, trend, slope, percentage (DQO Step 5).

- Limits on decision errors/precision, such as false acceptance error, false rejection error, power, and gray region (e.g., if $\alpha = 0.05$ and power $\geq 80\%$, 95% confidence interval must be 10% of mean) (DQO Step 6).

Constraints come in four categories: sampling/analysis, time/schedule, geographic, and budget. Sampling/analysis constraints could include instrumental performance (e.g., sensitivity and selectivity requirements for field or laboratory technologies), regulatory requirements specifying methodology, or weather constraints (e.g., performance of field technologies at low temperature, high humidity). Time/schedule constraints could include seasonal constraints such as the relationship of exposure to season (e.g., solvent volatility in warmer weather) and the availability of certified professionals. Geographic constraints could include physical barriers that may preclude sampling (e.g., rivers, fences) and also should include any possible hindrance to the ability to accurately identify sample location. Budget constraints are self-explanatory and determined in advance of the project. In addition, the design development should include existing regulations and requirements (e.g., state, municipal), data analysis needs, and secondary uses of the data.

What are the steps in the Sampling Design Process?

Steps of the sampling design process are represented in Figure 6 and described below.

Review the systematic planning outputs. First, the sampling objectives must be stated clearly. Next, make sure the acceptance or performance criteria are specified adequately (such as probability limits on decision errors or estimation intervals). Then review the constraints regarding schedule, funding, special equipment and facilities, and human resources.

Develop general sampling design alternatives. Decide whether the approach will involve episodic sampling events (where a sampling design is established and all data for that phase are collected according to that design) or an adaptive strategy (where a sampling protocol is established and sampling units are selected in the field, in accordance with the protocol,

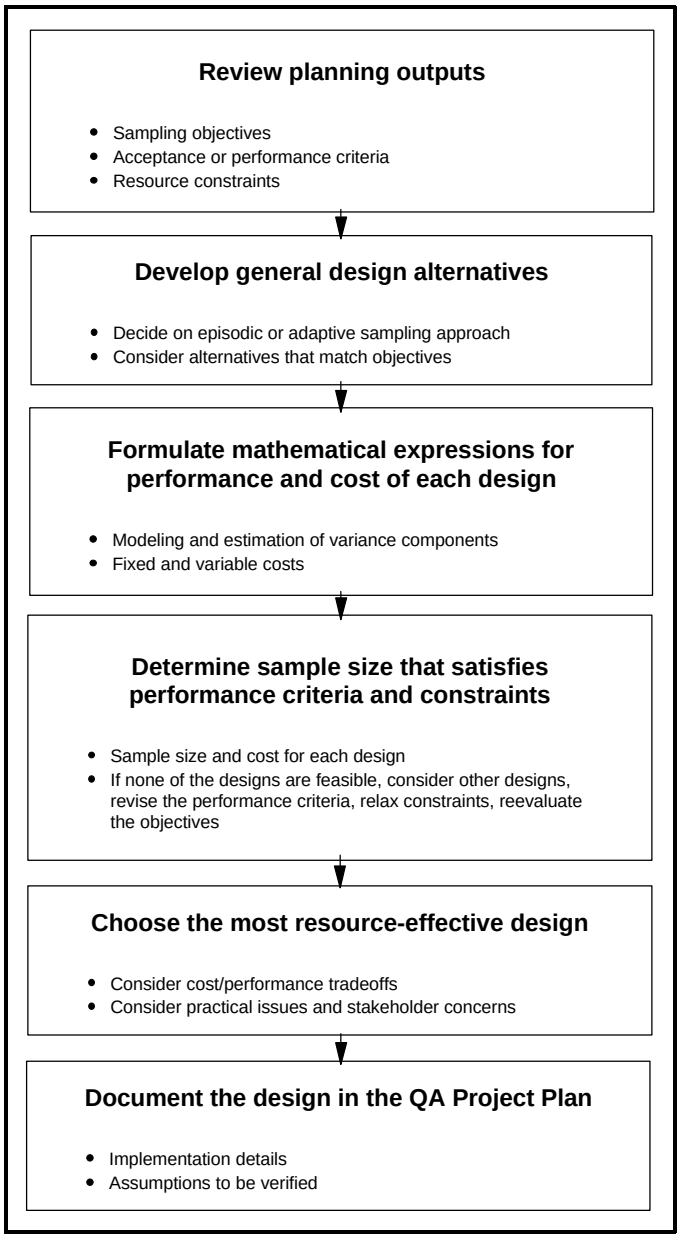


Figure 6. The Sampling Design Process

based on results from previous sampling for that phase). Consider sampling designs that are compatible with the sampling objectives. Evaluate advantages, disadvantages, and trade-offs in the context of the specific conditions of the study.

Formulate mathematical expressions for the performance and cost of each design alternative. For each design, develop the necessary statistical model or mathematical formulas needed to determine the performance of the design, in terms of the desired statistical power or width of the confidence interval. This process usually involves developing a model of relevant components of variance and estimating the total variance, plus key components as necessary. Also for each design, develop a cost model that addresses fixed costs (such as mobilization and setup costs) and variable costs (such as labor hours per sample and analytical costs per sample). Note that this step is not used in judgmental sampling designs.

Determine the sample size that satisfies the performance criteria and constraints. Calculate the optimal sample size (and sample allocation, for stratified designs or other more complex designs). Tradeoffs may be needed between less precise, less expensive measurement protocols (which allow for more sampling units to be selected and measured) and more precise, more expensive measurement protocols (which provide better characterization of each sampling unit at the expense of allowing fewer sampling units to be selected and measured). Verify that all constraints have been satisfied, including decision or estimation performance, schedule, cost, equipment, facilities, and personnel.

If none of the designs are feasible (i.e., performance requirements cannot be satisfied within all constraints), then consider the following corrective actions[2]:

- Consider other, perhaps more sophisticated, sampling designs.
- Relax the performance requirements (e.g., increase the allowable probability of committing a decision error).
- Increase the width of the gray region or estimation interval.
- Relax one or more constraints (e.g., increase the budget).
- Re-evaluate certain aspects of the sampling objectives (e.g., increase the scale of decision making, reduce the number of sub-populations that require separate estimates, or consider surrogate or indicator measurements).

Choose the most resource-effective design. Consider the advantages, disadvantages, and tradeoffs between performance and cost among designs that satisfy the performance requirements and constraints. Consider practical issues, schedule and budget risks, health and safety risks to project personnel and the community, and any other relevant issues of concern to stakeholders. Finally, obtain agreement within the planning team on the best design.

[2]Note that this step is not used in judgmental sampling designs.

Document the design in the QA Project Plan. Provide details on how the design should be implemented, contingency plans if unexpected conditions or events arise in the field, and quality assurance (QA) and quality control (QC) that must be performed to detect and correct problems and ensure defensible results. Specify the key assumptions underlying the design, particularly those that should be verified during implementation and assessment.

2.3 TYPES OF SAMPLING DESIGNS

Table 4 identifies the sampling designs discussed in this document, and indicates which chapter contained detailed information on each design. This section contains a brief description of each design, with some information regarding the type of applications for which each design is especially appropriate and useful.

Table 4. Sampling Designs Presented in this Guidance

Judgmental Sampling	Chapter 3
Simple Random Sampling	Chapter 4
Stratified Sampling	Chapter 5
Systematic and Grid Sampling	Chapter 6
Ranked Set Sampling	Chapter 7
Adaptive Cluster Sampling	Chapter 8
Composite Sampling	Chapter 9

2.3.1 Judgmental Sampling

In judgmental sampling, the selection of sampling units (i.e., the number and location of samples) is based on knowledge of the feature or condition under investigation and on professional judgment. Judgmental sampling is distinguished from probability-based sampling in that inferences are based on professional judgment, not statistical scientific theory. Therefore, conclusions about target population are limited and depend entirely on validity and accuracy of professional judgment, and probabilistic statements about parameters are not possible. Judgmental sampling may be used in conjunction with other sampling designs to produce effective sampling for defensible decisions.

2.3.2 Simple Random Sampling

In simple random sampling, individual units or locations are selected using random numbers, and all possible selections of a given number of units are equally likely. For example, a simple random sample of a set of drums can be taken by numbering all the drums and randomly selecting numbers from that list. This method is easy to understand, and the equations for determining sample size are relatively

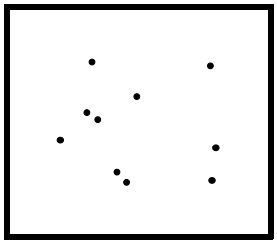


Figure 7. Simple Random Sampling

straightforward. An example is shown in Figure 7. Simple random sampling is useful when the population of interest is relatively homogeneous; i.e., there are not major patterns or "hot spots" expected. It is also most useful when the target population is relatively small. The main advantages of this design are that (1) it provides statistically unbiased estimates of the mean, proportions, and variability, (2) it is easy to understand and easy to implement, and (3) sample size calculations and data analysis are very straightforward.

2.3.3 Stratified Sampling

In stratified sampling, the target population is separated into non-overlapping strata, or sub-populations that are known or thought to be more homogeneous (relative to the environmental medium or the contaminant), so that there tends to be less variation among sampling units in the same stratum than among sampling units in different strata. Strata may be chosen on the basis of spatial or temporal proximity of the units or on the basis of pre-existing information or professional judgment about the site or process. Figure 8 depicts a site that was stratified on the basis of information about how the contaminant is present based on wind patterns and soil type and on the basis of surface-soil texture. This design is useful for estimating a parameter when the target population is heterogeneous and the area can be subdivided based on expected contamination levels. Advantages of this design are that is has potential for achieving greater precision in estimates of the mean and variance, and that it allows computation of reliable estimates for population subgroups of special interest.

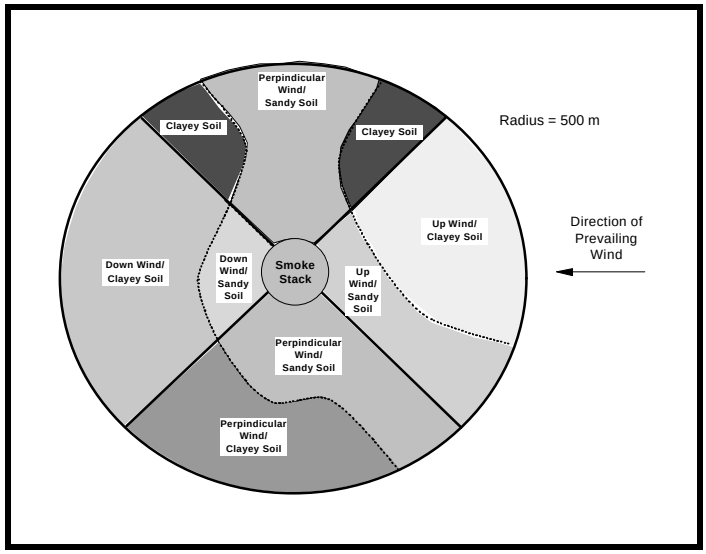


Figure 8. Stratified Sampling

2.3.4 Systematic and Grid Sampling

In systematic and grid sampling, samples are taken at regularly spaced intervals over space or time. An initial location or time is chosen at random, and then the remaining sampling locations are defined so that all locations are in regular intervals over an area (grid) or time (systematic). Examples of systematic grids include square, rectangular, triangular, or radial grids (Myers, 1997, pp. 445-446).

In random systematic sampling, an initial sampling location (or time) is chosen at random and the remaining sampling sites are specified so that they are located according to a regular

pattern (Cressie, 1993, p.317); for example, at the points identified by the intersection of each line in one of the grids shown in Figure 9. Systematic and grid sampling is used to search for hot spots and to infer means, percentiles, or other parameters and is also useful for estimating spatial patterns or trends over time. This design provides a practical and easy method for designating sample locations and ensures uniform coverage of site, unit, or process.

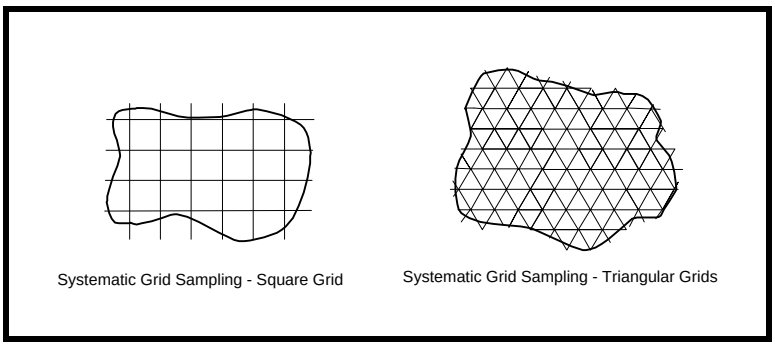


Figure 9. Systematic/Grid Sampling

2.3.5 Ranked Set Sampling

Ranked set sampling is an innovative design that can be highly useful and cost-efficient in obtaining better estimates of mean concentration levels in soil and other environmental media by explicitly incorporating the professional judgment of a field investigator or a field screening measurement method to pick specific sampling locations in the field. Ranked set sampling is a two-phase sampling design that identifies sets of field locations, utilizes inexpensive measurements to rank locations within each set, and selects one location from each set for sampling.

In ranked set sampling, *m* sets, each of size *r*, of field locations are identified using simple random sampling. The locations are ranked independently within each set using professional judgment or inexpensive, fast, or surrogate measurements. One sampling unit from each set is then selected for subsequent measurement using a more accurate and reliable (hence, more expensive) method for the contaminant of interest. Relative to simple random sampling, this design results in more representative samples and so leads to more precise estimates of the population parameters.

Ranked set sampling is useful when the cost of locating and ranking locations in the field is low compared to laboratory measurements. It is also appropriate when an inexpensive auxiliary variable (based on expert knowledge or measurement) is available to rank population units with respect to the variable of interest. To use this design effectively, it is important that the ranking method and analytical method are strongly correlated.

2.3.6 Adaptive Cluster Sampling

In adaptive cluster sampling, *n* samples are taken using simple random sampling, and additional samples are taken at locations where measurements exceed some threshold value. Several additional rounds of sampling and analysis may be required. Adaptive cluster sampling tracks the selection probabilities for later phases of sampling so that unbiased estimates can be

calculated despite over-sampling of certain areas. An example of an application of adaptive cluster sampling is delineating the borders of a plume.

Initial and final adaptive sampling designs are shown in Figure 10. Initial measurements are made of randomly selected primary sampling units using simple random sampling (designated by squares in Figure 10). Whenever a sampling unit is found to show a characteristic of interest (e.g., contaminant concentration of concern, ecological effect), additional sampling units adjacent to the original unit are selected, and measurements are made.

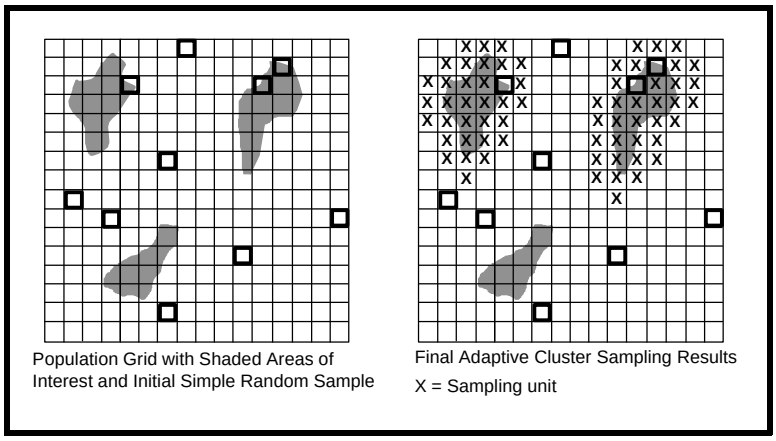


Figure 10. Adaptive Cluster Sampling

Adaptive sampling is useful for estimating or searching for rare characteristics in a population and is best suited for inexpensive, rapid measurements. It enables the delineations of the boundaries of hot spots, while also using all data collected with appropriate weighting to give unbiased estimates of the mean.

2.3.7 Composite Sampling

In composite sampling (shown in Figure 11), several volumes of material are physically combined and mixed in an effort to form a single homogeneous sample. This design can be cost-effective when analysis costs are large relative to sampling costs assuming there are no safety hazards or potential biases (e.g., loss of volatile organic components) associated with compositing.

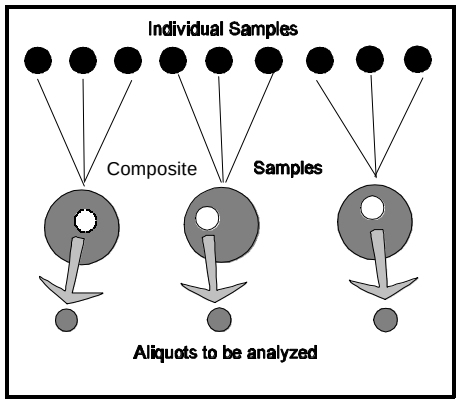


Figure 11. Composite Sampling

Compositing is used in conjunction with other sampling designs and when (1) the mean is the parameter of interest and (2) information on spatial or temporal variability is not needed. If individual samples can be retested, retesting schemes can be combined with composite sampling protocols to identify individual samples that have a certain trait, to recover some spatial or temporal information, and for application to air regulations when sampling is averaged over time—air is collected or passed through a filter over time ("averaged" means physical integration).

Compositing is useful when estimating a mean and when the cost of analysis is high compared to the cost of sample collection in the field. It is also useful when relatively large field variability obscures interpretation of data. It can be very cost-effective because it reduces the number of chemical analyses needed.

2.4 EXAMPLE SCENARIO

The following example illustrates the advantage of using the DQO Process for developing a sampling design. The example follows three sampling events: judgmental sampling (criterion for release decision), stratification and grid sampling within strata, and adaptive sampling.

At the company meeting, the project manager reports that, after sampling the site in areas where the most likely release would have occurred, the level of contamination in a few areas approached the action level for specific polynuclear aromatic hydrocarbons (PAHs) typically found in creosote. Based on the results of the judgmental sampling, they reached a consensus that the level of contamination should be characterized to a known level of confidence, and they decided to collect more data that would better represent the site. Using additional information about expected spatial patterns of contamination, they decide to stratify the site and use a grid design within strata for sampling. These data would then be used to estimate the mean concentration of the contaminant over the entire site.

At the next meeting, the project manager reports that the results from the grid sampling provided an adequate number of samples for derivation of the mean and confidence level. Therefore, she is now 95 percent confident that the overall concentration of PAHs at the site is below the action level. However, the manager also reports that the data indicated that the contaminant levels in one stratum were near the action level, so additional sampling may be needed to determine the extent of the problem. The team decides to use an adaptive sampling strategy to sample the stratum with contaminant levels above the action level, so that the portion of the stratum having the “hot spot” can be delineated and possibly remediated. The group then uses adaptive sampling to determine the boundaries of the hot spot. Once the size of the area is determined, the area is remediated. Follow-up sampling confirms that risks in this stratum are acceptable.

This example demonstrates, using the DQO Process, how different sampling designs can be efficiently coupled together to address evolving project objectives. Judgmental sampling was cost-effectively applied to assess whether or not any further study was warranted. Stratification coupled with grid sampling then resulted in a site-wide characterization of the discovered contamination problem to a desired level of confidence. And the project concluded by applying adaptive sampling at the one area to establish appropriate limits to the cleanup.

2.5 SELECTING A SAMPLING DESIGN

Table 5 presents examples of problem types that one may encounter and suggests sampling designs that are relevant for these problem types in particular situations.

Table 5. Choosing the Best Sampling Design for Your Problem

If you are...	and you have...	consider using...	in order to...
Performing a screening phase of an investigation and with an understanding of a relatively small-scale problem	a limited budget and/or a limited schedule	judgmental sampling	assess whether further investigation is warranted that should include a statistical probabilistic sampling design.
developing an understanding of when contamination is present	adequate budget for the number of samples needed	systematic sampling	have coverage of the time periods of interest.
developing an understanding of where contamination is present	adequate budget for the number of samples needed	grid sampling	have coverage of the area of concern and have a given level of confidence that you would have detected a hot spot of a given size.
estimating a population mean	adequate budget	systematic or grid sampling	also produce information on spatial or temporal patterns.
	budget constraints and analytical costs that are high compared to sampling costs	compositing	produce an equally precise or a more precise estimate of the mean with fewer analyses and lower cost.
	budget constraints and professional knowledge or inexpensive screening measurement that can assess the relative amounts of the contaminant at specific field sample locations	ranked set sampling	reduce the number of analyses needed for a given level of precision.
estimating a population mean or proportion	spatial or temporal information on contaminant patterns	stratified sampling	increase the precision of the estimate with the same number of samples, or achieve the same precision with fewer samples and lower cost.
delineating the boundaries of an area of contamination	a field screening method	adaptive sampling	simultaneously uses all observations in estimating the mean
estimating the prevalence of a rare trait	analytical costs that are high compared to sampling costs	random sampling and composite sampling	produce an equally precise or a more precise estimate of the prevalence with fewer analyses and lower cost.
assessing whether a population contains a rare trait	the ability to physically mix aliquot from the samples and then retest additional aliquots	composite sampling and re-testing	classify all samples at reduced cost by not analyzing every sample.

CHAPTER 3

JUDGMENTAL SAMPLING

3.1 OVERVIEW

Judgmental sampling refers to the selection of sample locations based on professional judgement alone, without any type of randomization. Judgmental sampling is useful when there is reliable historical and physical knowledge about a relatively small feature or condition. As discussed in *Quality Assurance Guidance for Conducting Brownfields Site Assessments* (U.S. EPA, 1998a), whether a judgmental or statistical (probability-based) sampling design should be employed is the main sampling design decision. This design decision applies not only to Brownfields investigations but also to most other environmental investigations. An important distinction between the two types of designs is that statistical sampling designs are usually required when the level of confidence needs to be quantified, and judgmental sampling designs are often required to meet schedule and budgetary constraints.

Implementation of a judgmental sampling design should not be confused with the application of professional judgment (or the use of professional knowledge of the study site or process). Professional judgement should *always* be used to develop an efficient sampling design, whether that design is judgmental or probability-based. In particular, when stratifying a population or site, exercising good professional judgement is essential so that the sampling design established for each stratum is efficient and meaningful.

3.2 UNDER WHAT CONDITIONS IS JUDGMENTAL SAMPLING APPROPRIATE?

For soil contamination investigations, judgmental sampling is appropriate for situations in which any of the following apply:

- Relatively small-scale features or conditions are under investigation.
- Small numbers of samples (e.g., fewer than 20 observations) will be selected for analysis/characterization.
- There is reliable historical and physical knowledge about the feature or condition under investigation.
- The objective of the investigation is to screen an area(s) for the presence or absence of contamination at levels of concern, such as risk-based screening levels (note that if such contamination is found, follow-up sampling is likely to involve one or more statistical designs).
- Schedule or budget limitations preclude the possibility of implementing a statistical design.

Judgmental sampling is often appropriate when addressing site-specific groundwater contamination issues. As further discussed in *Quality Assurance Guidance for Conducting Brownfields Site Assessments* (U.S. EPA, 1998a), a statistical sampling design may be unnecessary and impractical if data are needed to evaluate whether groundwater beneath a Brownfields site is contaminated because of the high cost of groundwater sample collection.

3.3 WHAT ARE THE BENEFITS OF USING JUDGMENTAL SAMPLING?

Because judgmental sampling designs often can be quickly implemented at a relatively low cost, the primary benefits of judgmental sampling are to meet schedule and budgetary constraints that cannot be met by implementing a statistical design. In many situations, when some or all of the conditions listed in Section 3.2 exist, judgmental sampling offers an additional important benefit of providing an appropriate level-of-effort for meeting investigation objectives without excessive consumption of project resources.

3.4 WHAT ARE THE LIMITATIONS OF USING JUDGMENTAL SAMPLING?

Judgmental sampling does not allow the level of confidence (uncertainty) of the investigation to be accurately quantified. In addition, judgmental sampling limits the inferences made to the units actually analyzed, and extrapolation from those units to the overall population from which the units were collected is subject to unknown selection bias.

3.5 HOW DO YOU IMPLEMENT JUDGMENTAL SAMPLING?

By definition, judgmental sampling is implemented in a manner left up to the professional(s) establishing the sampling design. Specialized academic and professional training is required before a professional is qualified to design a judgmental sampling program. The following paragraphs are only intended to provide a few examples of the most common factors that professionals should consider when establishing judgmental sampling designs.

As discussed in EPA's *Soil Screening Guidance* (U.S. EPA, 1996), current investigative techniques and statistical methods cannot accurately establish the mean concentration of subsurface soils within a contaminated source without a costly and intensive sampling program that is well beyond the level-of-effort generally appropriate for screening. The *Soil Screening Guidance* advises that, in establishing a judgmental sampling design to investigate subsurface soil contamination, the professional should locate two or three soil borings in the areas suspected of having the highest contaminant concentrations. If the mean contaminant concentration calculated for any individual boring exceeds the applicable numerical screening value, additional investigative phases should be conducted. The *Soil Screening Guidance* provides several calculation approaches for establishing a mean contaminant concentration for each boring; these approaches vary with the sampling-interval design.

In establishing a judgmental sampling design to investigate a subsurface soil contamination problem, the professional needs to consider many factors including the following:

- soil properties that affect contaminant migration (e.g., texture, layering, moisture content);
- the physical and chemical nature of the contaminant under investigation (e.g., solubility, volatility, reactivity);
- the manner in which the contaminant is understood to have been released (e.g., surface spill, leachate generated through above-ground or buried waste, leaking underground tank or pipe);
- the timing and duration of the release; and
- the amount of contaminant understood to have been released.

As stated in Section 3.2, judgmental sampling is often appropriate when addressing many site-specific groundwater contamination issues. A few of the most common factors to consider in establishing a judgmental sampling design to address a site-specific groundwater contamination issue include the following:

- the physical and chemical nature of the contaminant under investigation [e.g., solubility, volatility, reactivity, density (whether floating or sinking non-aqueous phase liquid could be present)];
- the possible effects of the migration of the contaminant through the unsaturated zone when and where the contaminant entered the aquifer;
- the possible ways the migration of the contaminant through the unsaturated zone might have changed the chemical nature of the contaminant before it entered the aquifer;
- the depths and thicknesses of aquifers beneath the site;
- the direction and rate of groundwater flow within each aquifer and variations in these parameters;
- the aquifer properties that cause the contaminant to disperse within the aquifer, both laterally and vertically; and
- the natural attenuation processes that may affect the contaminant as it migrates in groundwater.

3.6 UNDER WHAT CONDITIONS ARE OTHER DESIGNS COMMONLY USED IN CONJUNCTION WITH JUDGMENTAL SAMPLING?

There are two common situations when other sampling designs are used in conjunction with judgmental sampling. First, they may be used when the population or site is stratified, and judgmental sampling takes place within one or more strata. This situation is typical of small-scale soil contamination investigations when the suspected location of the contaminant release is known. The suspect area is identified as a stratum, and a judgmental sampling design is

established for that stratum. Other strata established for the site may be addressed through implementation of statistical sampling designs.

Second, other sampling designs may be used when judgmental sampling indicates that the screening criteria established for the area under investigation is exceeded, thereby warranting further investigation. Depending on how much historical information is available and how much information has been obtained from the judgmental-sampling phase, follow-up phases of investigation might involve any of the statistical sampling designs described in this guidance document.

3.7 EXAMPLES

3.7.1 Area Impacted by Contamination Can Be Visually Discerned

An active manufacturing facility is being sold, and the perspective purchaser is conducting an investigation to characterize existing environmental conditions and potential associated liability. One feature being assessed is a 5,000 ft^2 fenced area where drums of an aqueous cupric-chloride waste are stored. When released, the waste stains the soil blue-green. Eight blue-green stains are identified, and the typical stain size is about 1 yd^2. A judgmental sampling design is established whereby a single grab sample of soil is collected from each of the observed stains and analyzed for copper concentration. If any single copper result falls within one order of magnitude of the risk-based copper soil-screening level for industrial land use, the seller has agreed to pay for a follow-up investigation that will involve a statistical sampling program designed to better characterize the soil copper contamination and assess whether remediation is warranted.

3.7.2 Potential Location of the Contaminant Release Is Known

An abandoned textile mill is being investigated as a Brownfields site, and one previous employee was located who gave a reliable account of site features and activities. Based on this interview, the site was stratified and several different sampling designs (some statistical and some judgmental) were established. A judgmental sampling design is being used to investigate a 90-ft long drain pipe that carried a variety of wastes from one of the site factories to a leach field situated adjacent to the building; a statistical grid-sampling design was established to investigate the leach field. The drain pipe is accessible under a grating installed on the basement floor of the factory, and visual (external) and video (internal) inspections of the pipe showed it to be in good condition with no observable deterioration or cracks. However, several of the joints between the 10-ft length pipe segments appeared either loose or slightly separated. The judgmental sampling design established for this feature involved marking the basement floor adjacent to each pipe joint, removing the pipe, and collecting a single sample of the soil at each marked location for laboratory analysis. The analytical results then would be compared to the risk-based screening levels established for the list of potential site contaminants.

3.7.3 Contaminant Transport and Groundwater Flow Mechanisms Are Understood

State regulators have implemented a brief investigation to assess whether, over the 5 years of operation of a recently closed hog farm, releases from the onsite 0.25-acre animal waste lagoon have resulted in groundwater nitrate contamination. A publication issued by the U.S. Geological Survey (USGS) on the water resources of the county includes a detailed description of the aquifers within the area. Using the published information along with their geologic knowledge of the area, the state investigators knew that the aquifer was thick and laterally continuous, the water table is relatively shallow, and the regional groundwater flow direction is to the east. The available budget allowed the investigators to construct and sample only two wells on the farm property. Based on the USGS publication, which provided hydraulic conductivity values for the aquifer and a water-table map that indicated the hydraulic gradient, the investigators were able to calculate a typical groundwater flow velocity for the study area. The two new wells were constructed directly downgradient (due east) of the lagoons, one at a distance that represents a groundwater travel time of a few months and the other at a distance that represents a groundwater travel time of a few years. Using scientific understanding of the groundwater flow system and knowing that nitrate contamination would disperse laterally and vertically as it flows in groundwater, the investigators were confident that these wells were appropriately positioned to intercept a nitrate plume resulting from extensive lagoon leakage. For comparison to ambient (background) concentrations, the investigators had enough money in their budget to analyze samples from each of the five closest domestic wells installed in the same aquifer and situated upgradient (west) of the farm.

CHAPTER 4

SIMPLE RANDOM SAMPLING

4.1 OVERVIEW

Simple random sampling is the simplest and most fundamental probability-based sampling design. Most of the commonly used statistical analysis methods assume either implicitly or explicitly that the data are observations from a simple random sample.

A simple random sample of size *n* is defined as a sample selected from a population such that all possible samples of *n* elements have the same chance of being selected. Operationally, a simple random sample of size *n* occurs when *n* units are independently selected at random from the population of interest.

The most important characteristic of simple random sampling is that it protects against selection bias so long as the sampling frame (the list from which the sample is selected) is complete. Simple random sampling also is a benchmark against which the efficiency and cost of other sampling designs often are compared. Also, when using other sampling designs, the minimum sample size required often is estimated by computing the sample size that would be required with a simple random sampling design and multiplying that sample size by the relative efficiency of the alternative design (obtained from the literature or from other statistical considerations).

4.2 UNDER WHAT CONDITIONS IS SIMPLE RANDOM SAMPLING APPROPRIATE?

Simple random sampling may be appropriate if the population to be sampled is relatively small and homogeneous. However, simple random sampling usually is used in conjunction with other sampling designs, as discussed in Section 4.6.

Simple random sampling often is appropriate for the last stage of sampling when the sampling design has more than one stage of sampling (i.e., a sample of units is selected at the first stage and then sub-units are selected from each sample unit) (see Chapter 6 of Gilbert, 1987, and Chapters 12 and 13 of Thompson, 1992). Examples include the following:

- selecting one or more leaves for characterization from each sample plant,
- selecting one or more aliquots for chemical analyses from each soil sample, and
- assigning split samples or aliquots to laboratories or analytical methods.

In a similar vein, simple random sampling usually is required for assigning experimental units to treatments, or experimental conditions, in experimental designs.

4.3 WHAT ARE THE BENEFITS OF USING SIMPLE RANDOM SAMPLING?

The primary benefit of simple random sampling is that it protects against selection bias by guaranteeing selection of a sample that is representative of the sampling frame, provided only that the sample size is not extremely small (e.g., is 20 observations or more). Moreover, the procedures required to select a simple random sample are relatively simple.

Other benefits of using simple random sampling include the following:

- Statistical analysis of the data is relatively straightforward because most common statistical analysis procedures assume a simple random sampling design.

- Explicit formulas, as well as tables and charts in reference books, are available for estimating the minimum sample size required to support many statistical analyses.

4.4 WHAT ARE THE LIMITATIONS OF USING SIMPLE RANDOM SAMPLING?

Simple random sampling has two primary limitations:

- Because all possible samples are equally likely to be selected, by definition, the sample points could be geographically or temporally clustered by random chance. This limitation is somewhat overcome as the sample size increases, but it remains a limitation, even with large samples.

- Simple random sampling designs ignore all prior information, or professional knowledge, regarding the site or process being sampled. Prior information almost always can be used to develop a probability-based sampling design that is more efficient than simple random sampling (i.e., requires fewer observations to obtain the same level of precision).

4.5 HOW DO YOU IMPLEMENT SIMPLE RANDOM SAMPLING?

This section discusses how to determine the minimum sample size needed with simple random sampling to (1) estimate a population mean or proportion with pre-specified precision or (2) test a hypothesis regarding a population mean or proportion with pre-specified significance level and power. This section also addresses the process of selecting a simple random sample.

4.5.1 How do you estimate the sample size needed for Simple Random Sampling?

Determine the minimum sample size required to estimate a population proportion (e.g., proportion of units with concentrations above a health-based threshold) using a conservative preliminary estimate of the true population proportion. Use an estimate that is as close to 50 percent as realistically possible because using 50 percent results in the largest sample size (i.e., is

most conservative). In the absence of prior information, always use 50 percent as the preliminary estimate.

Determine the minimum sample size required to estimate a population mean (e.g., mean contaminant concentration) using a preliminary estimate of the population variance. The preliminary estimate should be large enough that the true population variance is not likely to be larger than the preliminary estimate because the sample size will be too small if the estimated variance is too small. Sources of a preliminary estimate of population variance include:

- a pilot study of the same population,
- another study conducted with a similar population, and
- an estimate based on a variance model combined with separate estimates for the individual variance components.

In the absence of prior information, estimate the standard deviation (square root of the variance) by dividing the expected range of the population by six, i.e.

$$\hat{\sigma} = \frac{\text{Expected Maximum} - \text{Expected Minimum}}{6}$$

However, this is only a crude approximation and should only be used as a last resort.

Appendix 4-A provides general-purpose formulas for determining the minimum sample size required to achieve specified precision for estimates of population means and proportions. Sample size formulas for achieving specified power for hypothesis tests are found in Section 3 of *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a). Appendix 4-B provides examples of the results from applying the formulas for determining the minimum sample size required to achieve sufficient precision.

4.5.2 How do you decide where to take samples with Simple Random Sampling?

Some environmental studies have distinct sampling units such as trees, fish, or drums of waste material. However, such distinct sampling units may not be available in environmental studies requiring samples of soil, water, or other solid or liquid media. In this case, the sampling units must be defined by the researcher and must be appropriate for selecting a representative sample of material from the medium of interest. The physical definition of a sampling unit in terms of its "size, shape, and orientation" is referred to as the sample "support" (Starks, 1986). Significantly, smaller support usually results in greater sampling variation (i.e., greater variability between sampling units) (see Section 21.5.3 of Pitard, 1993). For example, soil cores with a 2-inch diameter and 6-inch depth usually have greater variability in contaminant concentrations than cores with a 2-inch diameter and 5-foot depth, much like composite samples have less variability than individual specimens (see Chapter 9). Hence, the study objectives must clearly define the

sample support in order for the results (e.g., sample mean and variance) to be clearly interpretable.

Selecting a simple random sample is most straightforward when one can delineate a list of all the sampling units (e.g., barrels in a warehouse, trees at a study site). When selecting a simple random sample from a list of *N* distinct sampling units, follow this procedure:

- Label the population units from *1* to *N*.
- Use a table of random numbers to randomly select *n* integers from *1* to *N* from the list.

When *selecting a sample from a two-dimensional medium*, such as surface soils or the bottom of a lake or stream, the above one-dimensional list sampling approach can be used if an *M* by *N* grid is used to partition the population into *MN* unique units and the sample is selected from the list of *MN* units.

However, it is often more practical and flexible to select points directly at random in two-dimensional space if the desired sample support is not a rectangular area. If a rectangular coordinate system (i.e., (*x*, *y*)-coordinates, such as latitude and longitude) can be superimposed on the area of interest, then a simple random sample of points is generated by randomly generating *x*- and *y*-coordinates, as illustrated in Figure 12. Note that in an irregularly shaped sample area, randomly generated points falling outside of the sample area are not used. Alternatively, if the area of interest is a circular area, one can randomly generate values of a direction, θ, and a squared radius, r^2. Random values of the squared radius should be generated rather than the radius, *r*, because randomly generating radii, *r*, would over represent points near the center of the circle. Hence, if U_1 and U_2 are independent uniform random numbers on the interval (0,1) (e.g., four-digit decimal numbers from a table of random numbers), one would set $\theta = U_1$ and $r = \sqrt{U_2}$.

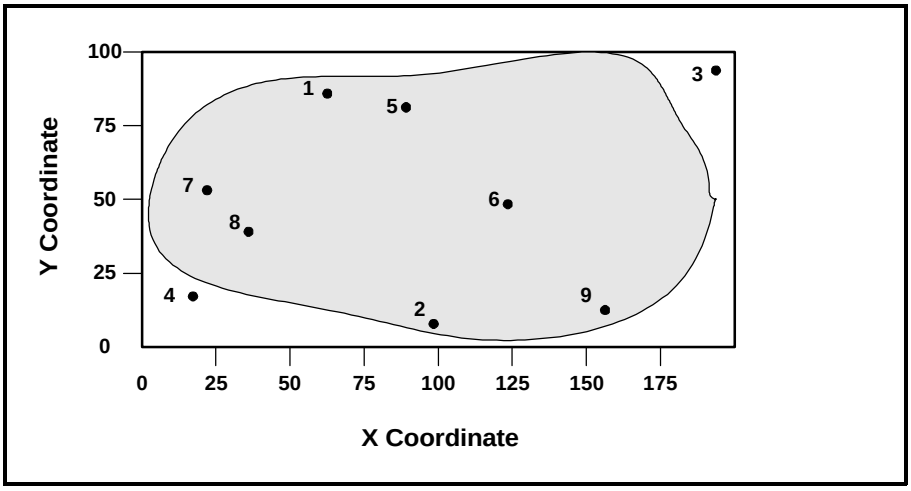


Figure 12. Example of a Map Showing Random Sampling Locations

When these sampling procedures are implemented to generate simple random samples in two dimensions, the randomly generated sampling points (i.e., *x*- and *y*- coordinates or direction, θ, and radius, *r*) should be rounded to the nearest unit that can be reliably identified in the field (e.g., nearest 1 or 5 meters). A sample specimen with the support defined in the sampling plan should then be obtained as near as possible to each of these approximate random sampling points using a procedure to avoid subjective bias factors such as "difficulty in collecting a sample, the presence of vegetation, or the color of the soil" (Section 5.3 of U.S. EPA, 1994). The protocols should be defined so that it will always be possible to obtain a sample from each randomly selected location. However, if it is physically impossible to obtain a specimen from a randomly selected location, deleting that location from the sample is valid so long as inferences are restricted to the accessible locations. If this restriction is untenable, the best compensation of loss of the observation is to take a sample from the most similar accessible location.

The above sampling methods can be extended fairly easily, at least conceptually, to *sampling three-dimensional wastes* (e.g., a waste pile or liquid wastes in a pond, lagoon, or drum). One approach is to superimpose a three-dimensional coordinate system over the area to be sampled [i.e., (*x*, *y*, *z*)-coordinates] and randomly generate *x*-, *y*-, and *z*-coordinates to identify randomly selected points. Alternatively, if the waste is contained in a cylinder or one can superimpose a cylinder over the area to be sampled, one can randomly generate direction, θ, squared radius, r^2, and height, *h*, to identify randomly selected points in three dimensions.

Although it is conceptually easy to generate random sampling points in three dimensions, actually getting a sampling tool into a three-dimensional medium at these randomly selected locations and extracting specimens with the correct sample support (size, shape, and orientation) can be difficult or impossible. Let us consider, for example, solid waste in a pile. If the waste pile has the consistency of soil, one may be able to take a core at the randomly selected location and extract a subsample from the core at the correct depth that has the desired support (e.g., 2-inch diameter and 6-inch depth). However, if the pile contains large impermeable solids (e.g., rocks of larger diameter than the core), taking such a core sample may not be possible. Alternatively, if the material is very fine, like ash, one may not be able to take a core sample because the process of getting the core would fundamentally alter the nature of the pile being sampled (e.g., it would cause the pile to shift or collapse). In that case, one potential solution may be to level the pile and take samples from the entire depth of the leveled pile at randomly selected points in two dimensions.

Liquid wastes present similar problems for sampling in three dimensions. If the liquid waste has the consistency of water, it may be possible to extract samples from randomly selected locations using a probe and pump. However, some wastes (e.g., a semi-liquid sludge) are too thick to be pumped yet not solid enough to extract competent cores. If one were sampling sludge from a lagoon, it might be necessary to sample entire vertical thickness of sludge at randomly selected locations (in two dimensions) and then analyze a subsample(s) from the resulting composite sample.

Pitard (1993) notes in Section 21.6.5 that one could theoretically obtain correct (representative) samples from a waste pile by selecting either one-dimensional or two-dimensional samples representing the full cross section of the waste. A one-dimensional sample is one in which vertical cross sections of a prescribed thickness are selected, as depicted in Figure 13. A two-dimensional sample is one in which cores from the top to the bottom of the waste pile are randomly extracted, as depicted in Figure 14. In Section 14.4.7, Pitard states that attempting to extract such samples is an "exercise in futility" because of the lack of appropriate sampling devices. Additional guidance regarding sampling devices and techniques that can be used to sample from three-dimensional waste piles is provided in Section 8.3 of Myers (1997).

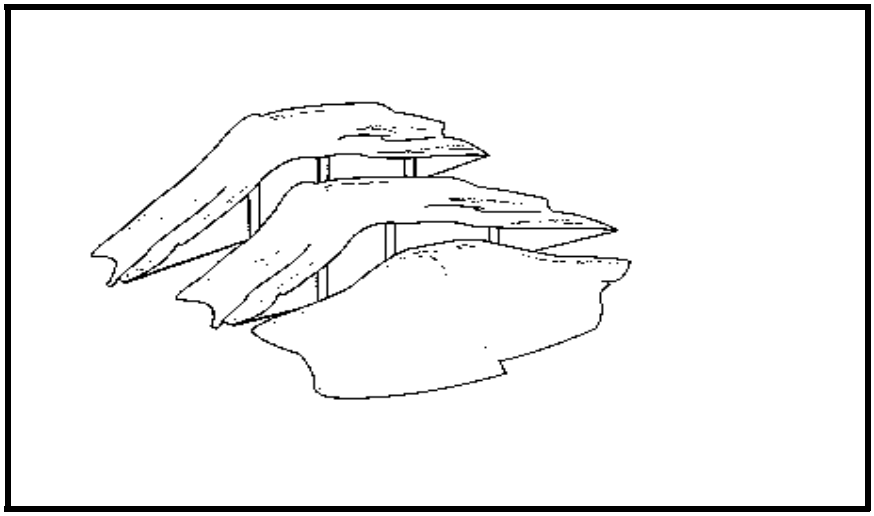


Figure 13. A One-Dimensional Sample of Cross-sections from a Waste Pile

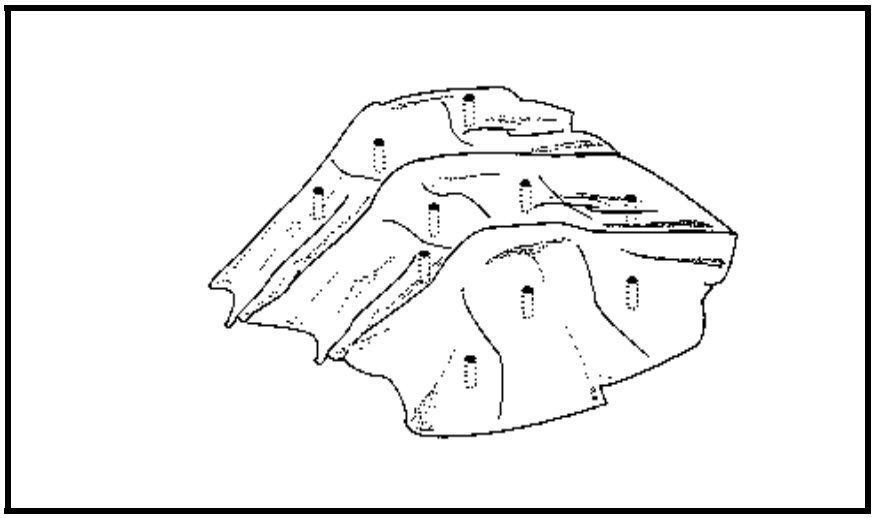


Figure 14. A Two-Dimensional Sample of Cores from a Waste Pile

An alternative sampling method that provides samples much like simple random samples is "quasi-random sampling." Quasi-random sampling refers to methods for generating a quasi-random sequence of numbers that are "in a precise sense, 'maximally avoiding' of each other" (Section 7.7 of Press et al., 1992). Samples in two or more dimensions are generated by pairing two or more of these quasi-random sequences. In two dimensions, the result is a set of sample points that, for any given sample size, appear to be uniformly scattered throughout the sampled area, as illustrated in Figure 15. Quasi-random sampling can be used to avoid the potential for geographic clustering that exists with simple random sampling without taking the risk of alignment of the sample with an unknown pattern of contamination, a limitation of grid sampling (as discussed in Chapter 6). The resulting data can be analyzed as if the sample were a simple random sample, knowing that the sampling variance is likely to be underestimated, which results in conservative inferences. Techniques for generating quasi-random samples are beyond the scope of this guidance but can be found in Section 7.7 of Press et al. (1992).

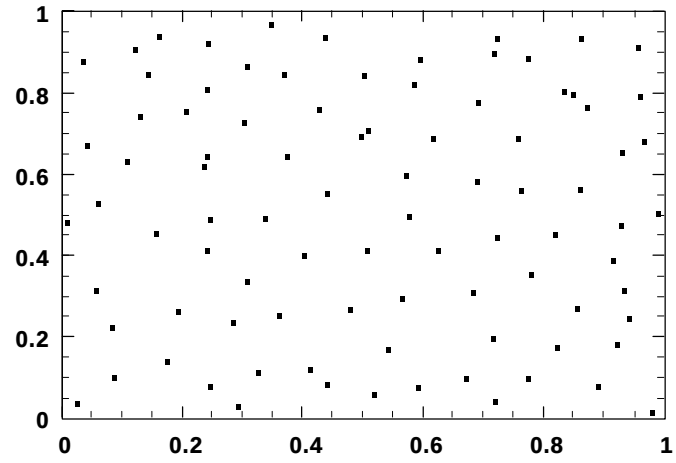


Figure 15. Illustration of a Quasi-Random Sample

4.6 UNDER WHAT CONDITIONS ARE OTHER DESIGNS COMMONLY USED IN CONJUNCTION WITH SIMPLE RANDOM SAMPLING?

Simple random sampling is used for selecting samples within sampling strata. When an independent simple random sample is selected from each stratum, the sampling design is referred to as stratified random sampling (described in Chapter 5). Simple random sampling is also used as the first step of the ranked set sampling process described in Chapter 6. It also can be used as the first step of the adaptive cluster sampling process described in Chapter 7.

4.7 EXAMPLES

4.7.1 General Simple Random Sampling Example

Suppose that a company with a fleet of 5,000 late-model, mid-sized sedans needs to determine an accurate estimate of the mean, or average, carbon monoxide (CO) emission rate of the fleet (in grams per mile, g/m). Suppose further that all vehicles in the fleet are late-model, 6-cylinder cars that are expected to have reasonably uniform emission rates. Hence, for selecting a sample of vehicles to be tested from this relatively homogeneous 5,000-vehicle population, a simple random sampling design is appropriate.

To determine the number of vehicles to be tested, the company researchers need an estimate of the variability to be expected in the CO emission rates. Given that the EPA standard for passenger cars is 3.4 g/m, they expect that the rate for most cars in the fleet will be between 1.0 and 3.0 g/m and that the standard deviation of measured CO emission rates will be no greater than about 0.50 g/m. Also, expecting a mean of about 2 g/m, they decide that a 95 percent confidence interval estimate with a half-width of no more than ± 0.10 g/m is desired. Hence, they apply the sample size formula found in Appendix Table 6 and find that the minimum number of vehicles to be tested is

$$n = \frac{\hat{\sigma}^2}{(d / z_{1-\alpha/2})^2} + 0.5\,(z_{1-\alpha/2})^2 = \frac{0.50^2}{(0.10 / 1.96)^2} + 0.5\,(1.96)^2 = 97.96$$

Because a fractional car cannot be sampled, the number of cars to be included in the sample was rounded up to 98.

The researchers assigned inventory control numbers to the cars in the fleet from 1 to 5,000, which were then used to facilitate the random sampling process. They used the random number table from an elementary statistics book to generate 98 random numbers between 1 and 5,000. The cars with these inventory control numbers were then selected as the simple random sample of cars to be tested for CO emission rates.

4.7.2 Example Using Look-up Tables in Appendix

Table 8: Alternatively, the researchers could arrive at the minimum sample size requirement calculated above by using Table 8 in Appendix 4-B. To use this table, they need to calculate

- CV = $\hat{\sigma}$ / μ = 0.50 / 2.0 = 25%
- Upper limit on the relative error = 0.10 / 2.0 = 5%

Table 8 shows that the minimum required sample size is 98 cars.

Table 9: Suppose the same company also wants to estimate the percentage of cars in the fleet with CO emission rates exceeding 3.0 g/m. If 10 percent or more cars exceed 3.0 g/m, they want the width of the 95 percent confidence interval estimate to be no more than ± 5 percent. Table 9 shows directly that at least 139 cars must be sampled to achieve this level of precision.

Table 10: Suppose the company decides that they need to overhaul the fleet to improve performance if the fleet's mean CO emission rate exceeds 2.5 g/m. To determine whether or not the overhaul is needed, they need to test the hypothesis

$$H_0: \mu \leq 2.5 \text{ g/m}$$
versus
$$H_A: \mu > 2.5 \text{ g/m}$$

In their application of the DQO process, the company officials determine that the maximum acceptable error rates are as follows:

- Type I: α = Prob(false acceptance when μ = 2.5 g/m) = 5%
- Type II: β = Prob(false rejection when μ = 2.75 g/m^3) = 5%

Table 10 then can be used to determine the minimum sample size needed by entering the table with the following parameters:

- α = Significance level = 5%
- Power = 1 - β = 95%
- Effect size = 100 * $|\mu_1 - \mu_0| / \hat{\sigma}$ = 100 * |2.75 - 2.50| / 0.50 = 50%

Table 10 indicates that a sample of 45 cars is necessary to achieve the error bounds specified for the hypothesis test.

Tables 11 and 12: Suppose the company decides that they need to overhaul the fleet of cars if more than 10 percent of the fleet have CO emission rates exceeding 3.0 g/m. To determine whether or not the overhaul is needed, they need to test the hypothesis

$$H_0: P \leq 10\%$$
versus
$$H_A: P > 10\%$$

In their application of the DQO process, the company officials determine that the maximum acceptable error rates are as follows:

- Type I: α = Prob(false acceptance when P = 10%) = 5%
- Type II: β = Prob(false rejection when P = 15%[1]) = 5%

Table 11 then can be used to determine the minimum sample size needed by entering the table with the following parameters:

- α = Significance level = 5%
- Power = 1 - β = 95%
- P_0 = 10%
- $|P_1 - P_0|$ = |15% - 10%| = 5%

[3] This number is chosen as the alternative hypothesis for which this level of β is to be achieved. For information on how to determine the appropriate numbers for any particular project, see Step 6 in the DQO Guidance [*Guidance for the Data Quality Objectives Process (EPA QA/G-4)* 1994].

Table 11 shows that a sample of 468 cars is necessary to achieve the error bounds specified for the hypothesis test.

Table 13: Suppose the company also has a fleet of 5,000 small pick-up trucks. The researchers want to know if the mean CO emission rate for their fleet of pick-up trucks exceeds that for the fleet of sedans. They then need to test the hypothesis

$$H_0: \mu_1 - \mu_2 \leq 0$$
versus
$$H_A: \mu_1 - \mu_2 > 0,$$

where μ_1 is the mean CO emission rate for the fleet of pick-up trucks and μ_2 is the mean CO emission rate for the fleet of sedans.

In their application of the DQO process, they determine that the maximum acceptable error rates are as follows:

- α = Prob(false acceptance when $\delta = \mu_1 - \mu_2 = 0$) = 5%
- β = Prob(false rejection when $\delta = \mu_1 - \mu_2 = 0.25$ g/m[1]) = 5%

Table 12 then can be used to determine the minimum sample size needed by entering the table with the following parameters:

- α = Significance level = 5%
- Power = 1 - β = 95%
- Effect size = $100 * |\delta_1 - \delta_0| / \hat{\sigma} = 100 * |0.25 - 0.00| / 0.50 = 50\%$

Table 13 shows that a sample of 88 sedans and 88 pick-up trucks is necessary to achieve the error bounds specified for the hypothesis test.

Tables 14 and 15: Suppose the company decides that they want to determine whether the proportion of pickup trucks in the fleet with CO emission rates greater than 3.0 g/m is greater than the proportion for the fleet of sedans. They then need to test the hypothesis

$$H_0: P_1 - P_2 \leq 0\%$$
versus
$$H_A: P_1 - P_2 > 0\%,$$

where P_1 is the proportion of pick-up trucks with emission rates exceeding 3.0 g/m and P_2 is the proportion of sedans with emission rates exceeding 3.0 g/m.

In their application of the DQO process, they determine that the maximum acceptable error rates are as follows:

- Type I: α = Prob(false acceptance) = 5%
- Type II: β = Prob(false rejection when P_1 = 10% and P_2 = 5%) = 5%

Table 14 then can be used to determine the sample size needed by entering the table with the following parameters:

- α = Significance level = 5%
- Power = 1 - β = 95%
- P_1 = 10%
- $|P_1 - P_{2|}$ = |10% - 5%| = 5%

Table 14 indicates that a sample of 947 sedans and a sample of 947 pick-up trucks are necessary to achieve the error bounds specified for the hypothesis test.

APPENDIX 4-A

FORMULAS FOR ESTIMATING SAMPLE SIZE REQUIREMENTS FOR SIMPLE RANDOM SAMPLING DESIGNS

This appendix is written primarily for an analyst with statistical training who is assisting with QA Project Plan development and review. It provides formulas for estimating the minimum sample size needed to achieve specified precision for estimates of means and proportions. It also provides advanced references to textbooks and literature articles that provide more in-depth information regarding sample size computations. The sample size formulas require a preliminary estimate of the variability of the observations, as discussed in Section 4.5.1.

Formulas for computing the minimum sample size required for simple random sampling are summarized in Table 6 for estimates of population means and Table 7 for estimates of population proportions. *In these tables, the expression z_α refers to the α-th percentile of the standard normal probability distribution*. Tables of the standard normal probability distribution are found in many statistics textbooks and are contained in *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a). To determine sample size requirements for hypothesis tests, one must first determine which hypothesis testing procedure will be applied and then determine the sample size requirements for that test. Guidance regarding how to select an appropriate hypothesis testing procedure and compute the minimum sample size required are provided in Section 3 of *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a). In every case, the calculated sample size should be rounded up to the next larger integer value (e.g., $n = 33.1$ rounds up to $n = 34$) because it is not possible to select a fractional portion of a sampling unit into the sample.

If the calculated sample size, n, is within 10 percent of the total number of units, N, in the population and all inferences will be restricted to only that population of N units, a "finite population correction" can be implemented to account for sampling a non-negligible portion of a finite population, and the minimum sample size can be reduced to the sample size, n', calculated as follows:

$$n' = \frac{n}{1 + n / N}$$

Table 6. Sample Size Requirements for Estimating Population Means

Estimation Problem	Sample Size Formula, Assumptions, and References
Using the mean, $\bar{x}$, from a simple random sample of n observations to estimate the population mean, μ	$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \, \hat{\sigma}^2}{d^2} + 0.5 \, z_{1-\alpha}^2$$ where α = false positive rate (Type I Error) β = false negative rate (Type II Error) $\hat{\sigma}^2$ = estimated variance of the population d = "within how much" of the true mean the sample mean must lie *Assumptions:* This method of calculating n is satisfactory providing n is large enough that the distribution of $\bar{x}$ is expected to be approximately a normal or Student's t probability distribution. *References:* Section 4.4 of Gilbert (1985) and Chapter 4 of Thompson (1992) cover these methods in the context of environmental studies. *Advanced References:* Sections 4.6 and 4.7 of Cochran (1977) and Section 5.2 of Desu and Raghavaro (1990) cover these methods in a more general context and provide a two-step method in which an initial sample from the population of interest is used to determine how many additional observations are needed. The second term of the sample size formula is based on Section 2 of Guenther (1981) and discussed in the appendix to EPA (1994a).

Table 7. Sample Size Requirements for Estimating Population Proportions

Estimation Problem	Sample Size Formula, Assumptions, and References
Using the sample proportion, *p*, from a simple random sample of *n* observations to estimate the population proportion, *P*	$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \, \hat{p} \, (1 - \hat{p})}{d^2} + 0.5 \, z_{1-\alpha}^2$$ where α = false positive rate (Type I Error) β = false negative rate (Type II Error) $\hat{\sigma}^2$ = estimated variance of the population d = "within how much" of the true mean the sample mean must lie *Assumptions:* This method of calculating *n* is satisfactory provided the calculated value of *n* is large enough that both *n*P > 10* and *n*(1 - P) > 10*. *References:* Section 5.3 of Thompson (1992) covers this method in the context of environmental studies. *Advanced References:* Section 4.4 of Cochran (1977) and Section 1.3.1 of Desu and Raghavaro (1990) present this method in a more general context; Section 5.2 of Desu and Raghavaro (1990) provides a two-step method in which an initial sample from the population of interest is used to determine now many additional observations are needed. The "Binomial Distribution" section in Kotz et al. (1982) provides an approximation based on the Poisson probability distribution that can be used when *P* is small (e.g., less than 0.1) and *n*P < 10*. Alternatively, when either *n*P < 10* or *n*(1 - P) < 10*, one can use the exact confidence limit charts in Appendix Table 4 of Conover (1980), based directly on the binomial distribution.

APPENDIX 4-B

SAMPLE SIZE TABLES FOR SIMPLE RANDOM SAMPLING DESIGNS

This appendix provides the following tables to determine the minimum sample size required to achieve sufficient precision with simple random sampling designs:

- Table 8. Sample Size Required for a 95% Confidence Interval Estimate of a Mean.
- Table 9. Sample Size Required for a 95% Confidence Interval Estimate of a Population Proportion.
- Table 10. Sample Size Required for a One-Sample t-Test.
- Table 11. Sample Size Required for a One-Sample Test for a Population Proportion, P, at a 5% Significance Level.
- Table 12. Sample Size Required for a One-Sample Test for a Population Proportion, P, at a 10% Significance Level.
- Table 13. Sample Size Required for a Two-Sample t-Test.
- Table 14. Sample Size Required for a Two-Sample Test for Proportions at a 5% Significance Level.
- Table 15. Sample Size Required for a Two Sample Test for Proportions at a 10% Significance Level.

The formulas that these sample size calculations are based upon are provided in the following documents:

- ***Appendix 4-A*** for the first two tables, which address sample size requirements for estimating means and proportions.
- Section 3 of *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a) for the remaining tables, which address sample size requirements for hypothesis tests.

Table 8. Sample Size Required for a 95% Confidence Interval Estimate of a Mean

Population	% Relative Error of the Estimated Mean					
CV	5%	10%	15%	25%	50%	75%
25%	98	26	13	6	3	3
50%	387	98	45	18	6	4
75%	867	219	98	37	11	6
100%	1539	387	173	64	18	9
125%	2403	603	269	98	26	13
150%	3460	867	387	141	37	18

The 95% confidence interval estimate of the population mean, μ, is given by

$$\bar{x} \pm d_r \mu \quad ,$$

where $\bar{x}$ is the sample mean and d_r is the maximum relative error, per Table 6.

Table 9. Sample Size Required for a 95% Confidence Interval Estimate of a Population Proportion

True Population Percentage	Maximum 95% Confidence Interval Half-Width					
	2%	3%	4%	5%	7.5%	10%
5%	457	203	NA	NA	NA	NA
10%	865	385	217	139	NA	NA
20%	1537	683	385	246	110	62
30%	2017	897	505	323	144	81
45%	2305	1025	577	369	164	93
50%	2402	1068	601	385	171	97

The 95% confidence interval estimate of the population percentage, *P*, is given by

$$p \pm d \quad ,$$

where *p* is the sample percentage and *d* is the maximum confidence interval half-width, per Table 7.

Table 10. Sample Size Required for One-Sample t-test

Significance Level	Power	Effect Size				
		10%	20%	30%	40%	50%
5%	95%	1,084	272	122	69	45
	90%	858	216	97	55	36
	80%	620	156	71	40	27
10%	95%	858	215	96	55	36
	90%	658	166	74	42	28
	80%	452	114	51	29	19

Per *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a), Section 3.2.1.1,
Case 1: H_0: $\mu \leq C$ vs H_A: $\mu > C$
Case 2: H_0: $\mu \geq C$ vs H_A: $\mu < C$.
In either case, the effect size is $100 * |\mu_1 - C| / \hat{\sigma}$, where $\mu = \mu_1$ is at the boundary of the gray region determined in Step 6 of the DQO process and $\hat{\sigma}$ is a preliminary estimate of the population standard deviation (square root of the variance).

Table 11. Sample Size Required for a One-Sample Test for a Population Proportion, P, at a 5% Significance Level

P_0		$\lvert P_1 - P_0 \rvert$			
Case 1	Case 2	5%	10%	15%	20%
	Significance level =		5%	Power =	95%
10%	90%	468	133	65	39
20%	80%	751	200	93	54
30%	70%	947	244	110	63
40%	60%	1056	266	118	65
50%	50%	1077	266	115	63
60%	40%	1012	244	103	54
70%	30%	860	200	80	39
80%	20%	621	133	46	NA
90%	10%	291	NA	NA	NA
	Significance level =		5%	Power =	90%
10%	90%	362	102	49	30
20%	80%	589	156	72	42
30%	70%	746	191	87	49
40%	60%	834	210	93	52
50%	50%	853	211	92	50
60%	40%	804	195	83	44
70%	30%	686	161	66	33
80%	20%	498	109	40	NA
90%	10%	239	NA	NA	NA
	Significance level =		5%	Power =	80%
10%	90%	253	69	33	20
20%	80%	419	109	50	29
30%	70%	534	136	62	35
40%	60%	600	151	67	38
50%	50%	617	153	67	37
60%	40%	583	142	61	33
70%	30%	501	119	50	26
80%	20%	368	83	32	NA
90%	10%	184	NA	NA	NA

Per *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a), Section 3.2.2.1:

Case 1: H_0: $P \leq P_0$ vs H_A: $P > P_0$

Case 2: H_0: $P \geq P_0$ vs H_A: $P < P_0$.

$P = P_1$ at the boundary of the gray region determined in Step 6 of the DQO process.

Table 12. Sample Size Required for a One-Sample Test for a Population Proportion, P, at a 10% Significance Level

P_0		$\lvert P_1 - P_0 \rvert$			
Case 1	Case 2	5%	10%	15%	20%
	Significance level =		10%	Power =	95%
10%	90%	378	109	54	33
20%	80%	601	161	75	44
30%	70%	753	195	88	50
40%	60%	837	211	93	52
50%	50%	852	210	91	49
60%	40%	798	191	80	42
70%	30%	676	156	62	30
80%	20%	484	102	34	NA
90%	10%	221	NA	NA	NA
	Significance level =		10%	Power =	90%
10%	90%	284	81	40	24
20%	80%	456	121	57	33
30%	70%	575	148	67	38
40%	60%	641	161	72	40
50%	50%	654	161	70	38
60%	40%	615	148	63	33
70%	30%	522	121	49	24
80%	20%	377	81	28	NA
90%	10%	177	NA	NA	NA
	Significance level =		10%	Power =	80%
10%	90%	188	53	25	15
20%	80%	308	81	38	22
30%	70%	392	100	45	26
40%	60%	439	110	49	28
50%	50%	449	111	49	27
60%	40%	424	103	44	24
70%	30%	363	86	36	18
80%	20%	265	59	22	NA
90%	10%	130	NA	NA	NA

Per *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a), Section 3.2.2.1:

Case 1: H_0: $P \leq P_0$ vs H_A: $P > P_0$

Case 2: H_0: $P \geq P_0$ vs H_A: $P < P_0$.

$P = P_1$ at the boundary of the gray region determined in Step 6 of the DQO process.

Table 13. Sample Size Required for a Two-Sample t-Test

Significance		Effect Size				
Level	**Power**	**10%**	**20%**	**30%**	**40%**	**50%**
5%	95%	2,166	542	242	136	88
	90%	1,714	429	191	108	70
	80%	1,238	310	139	78	51
10%	95%	1,714	429	191	108	69
	90%	1,315	329	147	83	53
	80%	902	226	101	57	37

Per *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a), Section 3.3.1.1,
Case 1: H_0: $\mu_1 - \mu_2 \leq \delta_0$ vs H_A: $\mu_1 - \mu_2 > \delta_0$
Case 2: H_0: $\mu_1 - \mu_2 \geq \delta_0$ vs H_A: $\mu_1 - \mu_2 < \delta_0$.
In either case, $\delta_1 = (\mu_1 - \mu_2)$ at the boundary of the gray region determined in Step 6 of the DQO process, and the effect size is

$$100 * |\delta_1 - \delta_0| / \hat{\sigma}.$$

See Table 2.4.1 in Cohen (1988) for a more extensive tabulation.

Table 14. Sample Size Required for a Two-Sample Test for Proportions at a 5% Significance Level

P_1		$\lvert P_1 - P_2 \rvert$			
Case 1	Case 2	5%	10%	15%	20%
	Significance level =		5%	Power =	95%
10%	90%	947	276	139	87
20%	80%	1510	406	192	114
30%	70%	1900	493	226	130
40%	60%	2116	536	240	136
50%	50%	2160	536	236	130
60%	40%	2030	493	212	114
70%	30%	1727	406	168	87
80%	20%	1250	276	106	NA
90%	10%	601	NA	NA	NA
	Significance level =		5%	Power =	90%
10%	90%	750	219	110	69
20%	80%	1195	322	152	90
30%	70%	1503	390	179	103
40%	60%	1675	424	190	108
50%	50%	1709	424	187	103
60%	40%	1606	390	167	90
70%	30%	1366	322	133	69
80%	20%	990	219	84	NA
90%	10%	476	NA	NA	NA
	Significance level =		5%	Power =	80%
10%	90%	541	158	80	50
20%	80%	863	232	110	65
30%	70%	1086	282	129	75
40%	60%	1209	307	138	78
50%	50%	1234	307	135	75
60%	40%	1160	282	121	65
70%	30%	987	232	96	50
80%	20%	715	158	61	NA
90%	10%	344	NA	NA	NA

Per *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a), Section 3.3.2.1:
Case 1: H_0: $P_1 - P_2 \leq 0$ vs H_A: $P_1 - P_2 > 0$
Case 2: H_0: $P_1 - P_2 \geq 0$ vs H_A: $P_1 - P_2 < 0$.

Table 15. Sample Size Required for a Two-Sample Test for Proportions at a 10% Significance Level

P_1		$\lvert P_1 - P_2 \rvert$			
Case 1	Case 2	5%	10%	15%	20%
	Significance level =		10%	Power =	95%
10%	90%	750	219	110	69
20%	80%	1195	322	152	90
30%	70%	1503	390	179	103
40%	60%	1675	424	190	108
50%	50%	1709	424	187	103
60%	40%	1606	390	167	90
70%	30%	1366	322	133	69
80%	20%	990	219	84	NA
90%	10%	476	NA	NA	NA
	Significance level =		10%	Power =	90%
10%	90%	575	168	85	53
20%	80%	917	247	117	69
30%	70%	1153	299	137	79
40%	60%	1285	326	146	83
50%	50%	1311	326	143	79
60%	40%	1232	299	129	69
70%	30%	1048	247	102	53
80%	20%	759	168	64	NA
90%	10%	365	NA	NA	NA
	Significance level =		10%	Power =	80%
10%	90%	395	115	58	37
20%	80%	629	170	80	48
30%	70%	792	206	94	55
40%	60%	882	224	100	57
50%	50%	900	224	98	55
60%	40%	846	206	88	48
70%	30%	720	170	70	37
80%	20%	521	115	44	NA
90%	10%	251	NA	NA	NA

Per *Guidance for Data Quality Assessment (QA/G-9)* (U.S. EPA, 1996a), Section 3.3.2.1:
Case 1: H_0: $P_1 - P_2 \leq 0$ vs H_A: $P_1 - P_2 > 0$
Case 2: H_0: $P_1 - P_2 \geq 0$ vs H_A: $P_1 - P_2 < 0$.

CHAPTER 5

STRATIFIED SAMPLING

5.1 OVERVIEW

Stratified sampling is a sampling design in which prior information about the population is used to determine groups (called strata) that are sampled independently. Each possible sampling unit or population member must belong to exactly one stratum. There can be no sampling units that do not belong to any of the strata and no sampling units that belong to more than one stratum. When the strata are appropriately constructed with respect to the variable being estimated, a stratified sampling design can produce estimates (mean, proportion, etc.) of the overall population with greater precision than simple random sampling. The use of proportional allocation to assign the samples to each stratum, regardless of how the strata are defined, will produce estimates of the population with greater precision at least as good as, and possibly better than, using a simple random sampling design. However, if optimal allocation is used to assign samples to the strata and the estimates of the variance within the strata are not close to the actual values, the level of precision in the resulting estimates could be worse than the level of precision for a simple random sample.

A stratified sampling design can also be used to obtain estimates for desired sub-populations or to ensure that important sub-populations have a sufficient number of sampling units in the samples. One of the most common uses of stratification is to account for spatial variability by defining geographic strata. Sampling by spatial strata may also be useful when study results need to be reported separately for particular geographic areas or regions. Strata may also be defined temporally. Temporal strata permit different samples to be selected for specified time periods and, hence, also permit the computation of separate time period level (e.g., seasonal) estimates having known precision. Temporal stratified sampling designs support accurate monitoring of trends.

5.2 UNDER WHAT CONDITION IS STRATIFIED SAMPLING APPROPRIATE?

The method of defining the strata depends on the purpose of the stratification. If estimates (mean, proportion, etc.) are desired for particular groups or regions, each group or region would be assigned as a stratum. If the investigator has prior knowledge of the spatial distribution of the study area, the strata should be defined so that the area within each stratum is as homogeneous as possible. In addition, the strata can be defined using reliable data on another variable that is highly correlated with the variable to be estimated. The resulting estimates will have greater precision than if no stratification were used. The variable providing the information used to establish the strata is referred to throughout this chapter as the “auxiliary variable.” Field conditions may require different sampling procedures to be used for different groups of the population in order to be efficient. This approach is facilitated by stratified sampling because each stratum can use a different sampling method.

5.3 WHAT ARE THE BENEFITS OF USING STRATIFIED SAMPLING?

For increased precision, the variable used to define the strata should be highly correlated with the response variable to be estimated. The amount of increase in precision over simple random sampling depends on the accuracy of the measurements of the second variable and the strength of correlation of the second variable with the variable to be estimated. Consider a situation in which a prior study had found that the amount of clay in the soil is correlated with the amount of a chemical that remains in the soil. In this case, the investigator could use a map of the study area showing the amount of clay in the soil to define the strata needed to estimate the concentration of the chemical. Strata can be defined in order to minimize costs to attain a given level of precision or to maximize precision for a given cost. Example 5-1 shows how the appropriate use of stratification in a planned sampling design can produce estimates with increased precision or require fewer samples as compared to simple random sampling.

Example 5-1. **Concentration of Arsenic in Surface Soil.** An investigator wants to estimate the average concentration of arsenic in the surface soil around the smoke stack at a hazardous waste incinerator facility to determine if the soil has been contaminated above the naturally occurring concentrations of arsenic for the region. Samples are to be taken within 500 meters from the smoke stack. Information gathered from prior studies indicates that the concentration of arsenic will be higher in the area along the prevailing wind direction and that the variability of the concentration of arsenic in the soil will be higher for clayey soils compared to sandy soils. Because the hazardous waste incinerator facility is located along the ocean coast, the prevailing winds flow both inland and offshore. The precision for the estimate of the concentration of arsenic can be increased by dividing the study area into strata according to the prevailing wind direction and the type of soil (see Figure 16).

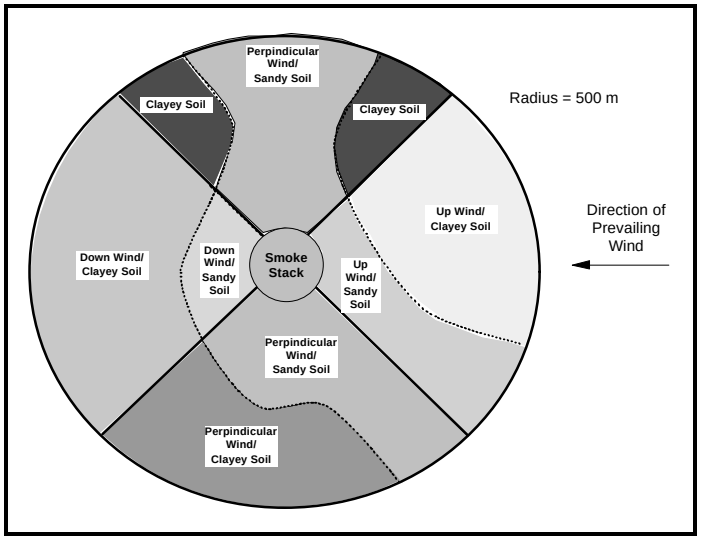


Figure 16. Stratification of Area to be Sampled

Budget restrictions will only allow 60 samples to be taken from the area around the smoke stack. The study area was stratified according to Figure 16, and the Neyman allocation (described in Section 5.5.1) was used to determine the number of samples to be randomly selected within each stratum. The summary statistics for the stratified samples are shown in Table 16. A simple random sample of 60 samples of soil was also taken from the study area. Table 16 shows that taking 60 samples by simple random sampling and stratified random sampling produce similar estimates for the mean concentration of arsenic, but the standard error associated with the stratified random sample is much lower (i.e., the precision is much higher) than that of the simple

random sample. Table 17 shows that the investigator would have only needed to take 8 soil samples using stratified random sampling in order to get a precision similar to that obtained by analysis of 60 samples taken by simple random sampling. This result is shown by comparing the standard errors and the 95 percent confidence intervals shown for the various samples sizes under stratified random sampling and simple random sampling. If a particular precision was desired for this study (e.g. a standard error of 1.00 for estimating the mean), the investigator could reduce the costs of obtaining an estimate of the average concentration of arsenic by using a stratified sampling design as described above instead of a simple random sampling design.

Table 16. Summary Statistics for Simple and Stratified Random Samples

	Simple Random Sampling	Stratified Random Sampling				
		Down Wind/ Clayey Soil	Don Wind/ Sandy Soil	Perpendicular Wind/Clayey Soil	Perpendicular Wind/Sandy Soil	Overall
# samples	60	43	5	10	2	60
mean	19.81	46.16	12.66	9.49	10.20	22.94
standard error	4.35	9.99	4.63	2.28	3.12	0.48

Table 17. Number of Samples Needed to Produce Various Levels of Precision for the Mean

	Simple Random Sampling	Stratified Random Sampling					
# samples	60	60	40	20	14	9	8	7
standard error	4.35	0.48	0.71	1.43	2.02	3.02	3.44	4.00
95% CI	±8.69	±0.95	±1.44	±2.99	±4.37	±6.83	±7.93	±9.47

Stratification can be useful when the implementation of different sampling designs in each stratum could reduce costs associated with the sample selection. The strata can be defined in order to minimize costs associated with sampling at various sites. Study sites that are close in proximity to one another can be assigned to one stratum to minimize the travel time for a team of field personnel to take samples at these locations. Also, if the costs of collecting samples at a portion of a study site are much greater than the rest of the study site, the most costly portion of

the site can be assigned as a stratum to minimize sample collection costs. Groups of the population with certain characteristics, which may or may not be the same as the primary stratification variables, can be used as strata in order to ensure that a sufficient number of sampling units appear in the sample for estimates or other analysis of the groups. For example, the investigator may want to stratify the country by average yearly rainfall in order to increase the precision of estimates and also stratify by EPA region to obtain estimates for each region. Stratification can also ensure that certain rare groups of the population that are of interest for estimates or analysis, and that may not otherwise have sufficient sample sizes, have the sample sizes necessary to perform the desired analyses.

5.4 WHAT ARE THE LIMITATIONS OF USING STRATIFIED SAMPLING?

Stratified sampling requires reliable prior knowledge of the population in order to effectively and appropriately define the strata and allocate the sample sizes. The gains in the precision or the reductions in cost depend on the quality of the information used to set up the stratified sampling design plan. Any possible increases in precision are particularly dependent on strength of the correlation of the stratification variable with the variable to be estimated. Precision may be reduced if Neyman or optimal allocation is used and the auxiliary data used for the optimization calculations are not applicable for the current study site.

After the optimal stratified sampling plan is completed, the investigator may encounter difficulties selecting the samples in the field. The investigator's ability to gain access to all sections or strata to be sampled may be limited by various geographic, physical, or meteorological conditions. Such limitations could reduce the gains in precision that were expected for the results.

The advantages and disadvantages of stratified sampling relative to other designs that use additional data sources, such as ranked set sampling and grid sampling, are compared in the last chapter of this document.

5.5 HOW DO YOU IMPLEMENT STRATIFIED SAMPLING?

5.5.1 How do you decide what sample size to use with this design?

Unless budget restrictions limit the total number of samples that can be selected, the strata should be determined before allocating the sample sizes. As mentioned earlier, the methods used to define the strata depend on the reasons that stratification is desired. When the strata are to be defined according to a second variable that is correlated with the variable to be estimated, the optimal definition of the strata is to allocate the strata so that the population included in each stratum is as homogeneous as possible with respect to the second variable.

Cochran (1977) offers some guidelines on how to optimally assign strata when the second variable is continuous (i.e., contains measurement values). If the investigator is interested in estimating the overall mean for the population, Cochran suggests defining no more than six strata

and using a procedure attributed to Dalenius & Hodges in order to determine the optimal cutoff values for each of the strata based on the distribution of the second variable for the population. The steps for determining the Dalenius-Hodges strata are given in Section 5A.6 of Cochran (1977).

Once the strata have been defined, a number of options can be used to allocate the sample sizes to each stratum. Equal allocation can be used to assign the same number of samples to be selected within each stratum. Proportional allocation can be used to allocate the samples to the strata so that the proportion of the total sampling units allocated to a stratum is the same as the proportion of sampling units in the population that are classified in that stratum. As mentioned in Section 5.1, proportional allocation can ensure that the precision of the population estimates will be as least as good as, if not better than, the precision without the use of stratification. Optimal allocation has two options:

- Optimize the precision for a fixed study cost.
- Optimize the cost of study for a fixed level of precision.

If the investigator has a fixed budget in order to collect the samples, the samples could be allocated so that the results would produce the highest precision for the variable to be estimated. If the investigator is required to obtain a specified level of precision, the samples could be allocated so that the costs in obtaining the designated level of precision are as low as possible. A special case of the optimal allocation in which the cost of sampling each unit is the same across all strata is Neyman allocation. As previously stated, the benefits of the stratified sampling design, especially when the optimal sample allocations are used, depend on the quality of the data used to set up the sampling design and the strength of the correlation between the auxiliary variable and the variable to be estimated. However, because the optimal and Neyman sample allocations depend on auxiliary data, the increase (or possible decrease) in precision of the estimates as compared to simple random sampling depends on the accuracy of the variance values used in the sample allocation calculations. The formulas for the sample size allocations can be found in Appendix 4A.

5.5.2 How do you decide where to take samples with this design?

Any sampling design can be used to select the samples within each stratum. Where to select these samples will depend on the choice of sampling design that is used (Section 5.6.).

5.6 UNDER WHAT CONDITIONS ARE OTHER DESIGNS COMMONLY USED IN CONJUNCTION WITH STRATIFIED SAMPLING?

As mentioned earlier, any sampling design can be used within each stratum. The choices include, but are not limited to, simple random sampling, quasi-random sampling, grid sampling, and even another level of stratified sampling.

APPENDIX 5-A

Notation:

L	number of strata
N_h	total number of units in stratum h
N	total number of units in population, $N = \sum_{h=1}^{L} N_h$
n_h	number of units sampled in stratum h
n	total number of units sampled, $n = \sum_{h=1}^{L} n_h$
σ_h	prior known standard deviation in stratum h
W_h	stratum weight, $W_h = N_h / N$
C	total budget
C_0	initial fixed costs
C_h	cost per sample for stratum h
V	fixed variance

- To calculate the mean and standard deviation for stratified sampling:

$$\text{mean: } \bar{y}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h$$

$$\text{st dev: } sd(\bar{y}_{st}) = \left(\frac{1}{N^2} \sum_{h=1}^{L} N_h^2 (1 - n_h / N_h) \frac{\hat{var}(y_h)}{n_h} \right)^{1/2}$$

- To calculate the sample size within the stratum:

  S equal allocation

$$n_h = \frac{n}{L}$$

  S proportional allocation

$$n_h = n \frac{N_h}{N}$$

  S Neyman allocation

$$n_h = n \frac{N_h s_h}{\sum_{k=1}^{L} N_k s_k}$$

- optimal allocation for fixed cost

$$n_h = \frac{(C - C_0) N_h s_h / \sqrt{C_h}}{\sum_{k=1}^{L} N_k s_k \sqrt{C_k}}$$

- optimal allocation for fixed precision

$$n = \frac{\left(\sum_{h=1}^{L} W_h s_h \sqrt{C_h} \right) \sum_{h=1}^{L} W_h s_h / \sqrt{C_h}}{V + \frac{1}{N} \sum_{h=1}^{L} W_h s_h^2}$$

CHAPTER 6

SYSTEMATIC/GRID SAMPLING

6.1 OVERVIEW

Systematic sampling, also called grid sampling or regular sampling, are based on a specified pattern and have samples taken at regular intervals along that defined pattern. Systematic designs are good for uniform coverage, ease of use, and the intuitive appeal that nothing is missed. Samples taken at regular intervals, such as at every node of an area defined by a grid, are useful when the goal is to estimate spatial or temporal correlations or to identify a pattern.

Systematic sampling is used to ensure that the target population is fully and uniformly represented in the sample. The regular assignment of locations or events to the sample provides assurance that the sample truly represents the overall characteristics of the target population. To make systematic sampling a probability-based design, the initial unit for the first sample of a sample of size *n* is chosen at random; then the remaining (*n*-1) units are chosen so all *n* are located according to some pattern.

There are two major applications for systematic sampling:

- **Spatial designs.** Samples may be selected in one, two, or three dimensions if the population characteristic of interest has a spatial component. Sampling along a line or transect represents sampling in one dimension. Sampling every node on a grid laid over an area of interest is sampling in two dimensions. If depth or volume is of interest, samples can be taken at regular intervals in three dimensions, such as uniformly sampling a pile of dirt. Several options for patterns for two-dimensional sampling in space are shown in Figure 17. In Figures 17 (a), (b), and (d), node A is randomly assigned and all other nodes follow according to A along a specified pattern. In Figure 17 (c) all nodes are randomly assigned.

- **Temporal and/or periodic designs.** When samples are selected to represent a target population that changes over time, we are back to a one-dimensional problem where every kth unit is selected, or a sample is collected every 10 minutes. Figure 18 shows an example of periodic sampling.

Systematic sampling designs are used in three situations:

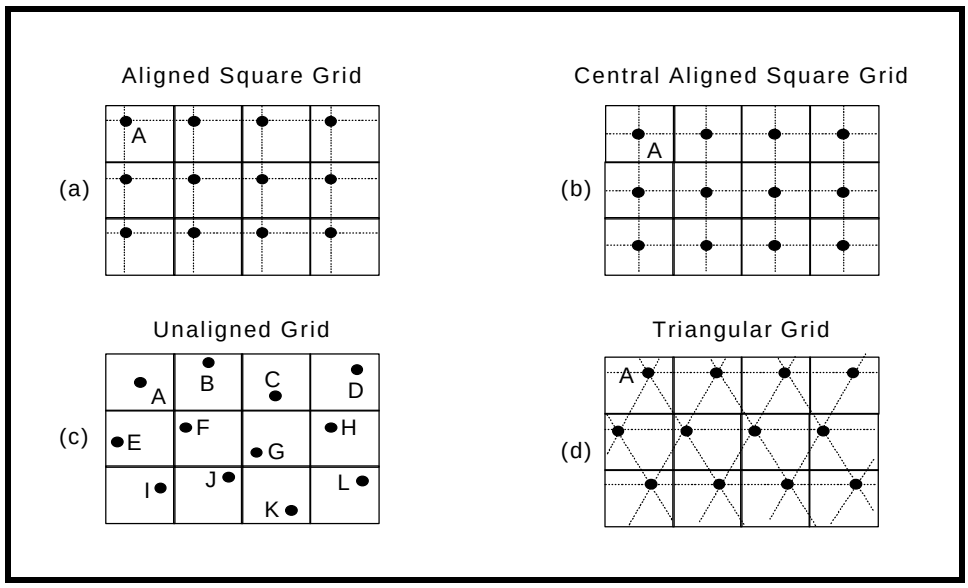


Figure 17. Systematic Designs for Sampling in Space (Gilbert, 1987, p. 94)

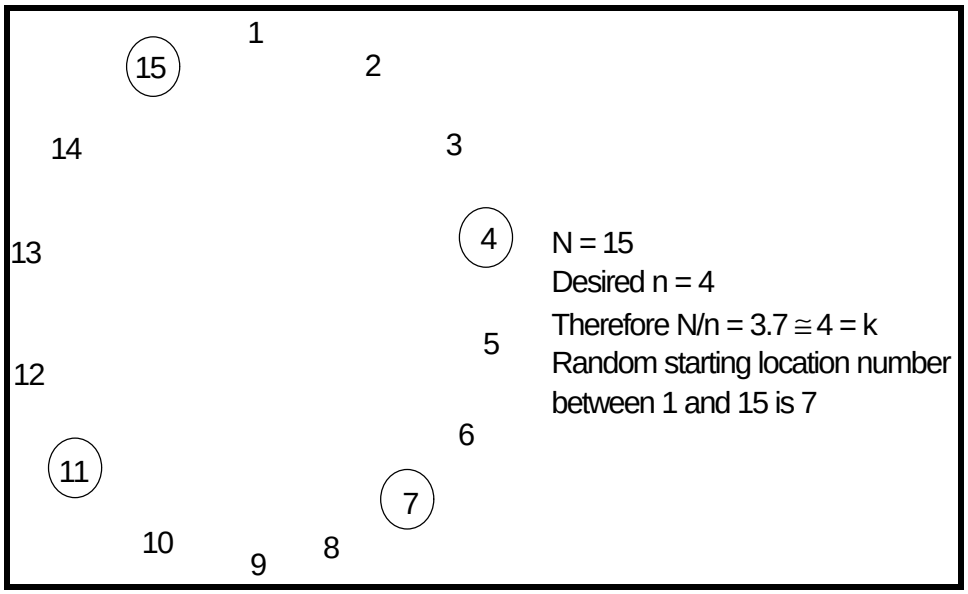


Figure 18. Choosing a Systematic Sample with Prespecified Number of Samples n, and Known Value of Number of Population Units N (Gilbert, 1987, pp. 91-92)

1. When making an inference on a population parameter, such as a mean about environmental measurements that are known to be heterogeneous. A systematic design is only one of many sampling designs that may be used for making an inference of a population parameter. However, if the units in the target population are spatially correlated and, hence, data are clumped as opposed to showing a more homogenous pattern of contamination, then systematic sampling will often be more efficient (more precise for a given amount of sampling) than random sampling at estimating quantities such as the overall mean or the total.

2. When estimating a trend or identifying a spatial or temporal correlation. The systematic design is uniquely suited for this type of problem because a known and regular distance or time interval between samples allows an association between increasing or decreasing data values and fixed increments of time or space. Patterns can be identified because nodes that register a "hit" can be connected to provide an estimate of the shape of the item of interest. Random sampling to achieve similar information would require more samples.

3. When looking for a "hot spot" or making a statement about the maximum size object that could be missed with a given sampling design. This is a geometry problem. An object of a specified size or larger within a grid of a certain spacing cannot be used without at least one of the grid nodes falling on the object (i.e., getting a "hit"). The regularity and complete coverage of nodes permits statements such as "there is a 95 percent confidence that no 55-gallon drum exists in this area" based on sampling results.

Grid designs can have various shapes, orientations, and selection criteria for the initial grid node. This flexibility, the intuitive appeal, and easily explained protocol for taking regular samples make systematic sampling one of the more popular and defensible sampling designs.

6.2 UNDER WHAT CONDITIONS IS SYSTEMATIC/GRID SAMPLING APPROPRIATE?

Systematic sampling is often used in environmental applications because it is practical and convenient to implement in the field. It often provides better precision (i.e., smaller confidence intervals, smaller standard errors of population estimates) and more complete coverage than unrestricted random sampling. Systematic sampling is appropriate if either of the following conditions pertain:

- There is no information about a population and the objective is to determine if there is a pattern or correlation among units, or
- There is a suspected or known pattern or correlation among units at the site and the objective is to estimate the shape of the pattern or the strength of the correlation.

If distinct features exist at a site, such as an ecological cluster or a groundwater plume, then collecting data on a regular grid is the most efficient approach to ensuring such features are actually sampled with a high probability.

Systematic sampling would be inappropriate if a known pattern of contamination coincides with the regularity of the grid design. Such a coincidence would result in an overestimation or underestimation of a particular trait in the target population of interest. For example, if a line of trees resulted in soil mounds with high contamination along the treeline and a grid line was aligned with the tree line, a decision about the average contamination over an entire area would be upwardly biased by so many samples collected in the high concentration area along the tree line. If any prior information is available on the possible patterns of contamination, this information may be important in selecting grid spacing, grid orientation, and whether or not systematic sampling designs have an advantage over other designs.

Cochran (1977) finds that systematic sampling can be considerably more precise than simple random sampling or even stratified random in some situations. He states: "Systematic sampling is more precise than simple random sampling if the variance within the systematic samples is larger than the population variance as a whole. Systematic sampling is precise when units within the same sample are heterogeneous and is imprecise when they are homogeneous" (p. 208). Cochran demonstrates that systematic sampling is capable of providing enhanced performance over other designs depending on the properties of the target population. He reports results from a study of 13 different data sets from natural populations showing a consistent gain in precision using systematic sampling.

Gilbert (1987, p. 95) finds a similar dependency of relative performance of systematic sampling based on trends, patterns, and correlations present in the target population. He compares systematic sampling for four types of population structures:

- Populations in random order
- Populations with linear trends
- Populations with periodicities
- Populations with correlations between values in close proximity

Gilbert makes three conclusions. First, for populations in random order, we might as well use systematic sampling because of the convenience. Populations in random order would be a reasonable scenario for contamination deposited by wind, erosion over time, and multiple processes such as drums moved many times over the years and any situation where no prior information exists. Second, if the population consists entirely of a linear trend, systematic sampling will, on the average, give a smaller Var $\bar{x}$ (i.e., sampling error of the sample mean) than simple random sampling. Third, stratified random sampling will, on average, give a smaller Var $\bar{x}$ than either systematic sampling or simple random sampling. However, the performance of systematic sampling for estimating a mean can be improved by using a weighted estimate of the mean.

A comprehensive study by Yfantis, Flatman, and Behar (1987) discusses the level of efficiency and accuracy of different grid types. They conclude that an equilateral triangular grid works slightly better for the majority of the cases they studied. However, this study did not include the effects of a second or additional phases of sampling. It is possible that when a multiple-phase sample is planned, the specific type of first-phase grid may be less important than using geostatistical techniques (such as geostatistical simulations) to locate second-phase samples in locations that most reduce probabilities of estimation errors (EPA, 1996, p. 6-3).

6.2.1 Soil Contamination Applications

For applications where the goal of sampling is to evaluate the attainment of cleanup standards for soil and solid media, EPA guidance (EPA, 1992) recommends collecting samples in the reference areas and cleanup units on a random-start equilateral triangular grid except when the remedial-action method may leave contamination in a pattern that could be missed by a triangular grid; in this case, unaligned grid sampling is recommended.

6.2.2 Ecological and Environmental Survey Applications

The National Stream Survey and EPA's Environmental Monitoring and Assessment Program are two large-scale environmental surveys that use variable probability, systematic sampling and a special estimator called the Horvitz-Thompson estimator to estimate population parameters of ecological interest. For the National Stream Survey, all streams represented as blue lines on 1:250,000 topographic maps define the target population of streams. Sampling units were selected using a square grid, with density of 1 grid node per 64 square miles, imposed on 1:250,000 topographic maps of a target area. A target stream reach was selected into the sample if a grid node fell into the direct watershed of that reach. This protocol resulted in reaches being sampled with probability proportional to direct watershed area. In the Environmental Monitoring and Assessment Program, one objective is to estimate the current condition of the nation's ecological resources on a regional basis with known confidence. The Environmental Monitoring and Assessment Program's sampling design is based on a systematic, triangular grid (also see discussion in Section 6.6). The grid is used to select a sample in a manner analogous to the National Stream Survey. For example, for sampling lakes, each lake is identified by its "center" and a grid node identifies a lake to be included in the sample as the lake that has a center closest to the grid node. The probability of sampling a given lake is proportional to the area of the polygon enclosing the region closer to that lake's center than to any other lake's center. Larger lakes have a higher probability of being included in the sample (Stehman and Overton, 1994, p.2).

When estimating abundance for various animals, samples are often taken along a transect at regular intervals. This is a form of grid sampling. A pronghorn (antelope) abundance study evaluated the efficiency of systematic sampling versus simple random sampling versus probability proportional to size sampling (Kraft et al., 1995, pp.130-131). The total number of pronghorn was already known; this was a simulation study to evaluate alternative sampling plans. The sampling unit was a 0.8-km-side linear transect variable in length according to size and shape of

the study area. Six different study areas were used. A plane flew along the transect and when a pronghorn was sighted, the pilot circled until the herd could be counted. The goal was to estimate total abundance of pronghorn in an area. For the systematic sampling, the sampling units (transects of different lengths) in an area were numbered; after the first unit was randomly chosen, every *p*th unit following was selected. For this study, it was found that stratification combined with accurate estimates of optimal stratum sample sizes increased precision, reducing the mean coefficient of variation from 33 without stratification to 25 with stratification. Cost, however, increased with stratification by 23 percent.

6.2.3 Groundwater Applications

For sampling groundwater in fixed wells over time, a systematic sample in time is usually preferred over a simple random sample in time. There are several reasons for this preference: extrapolating from the sample period to future periods is easier with a systematic sample than a simple random sample; seasonal cycles can be easily identified and accounted for in the data analysis; a systematic sample will be easier to administer because of the fixed schedule for sampling times; and most groundwater samples have been traditionally collected using a systematic sample, making comparisons to background more straightforward.

EPA guidance on groundwater sampling for evaluating attainment of cleanup standards (EPA,1992) suggests a variation of systematic sampling when periodic seasonal variations or other repeated patterns are suspected. Several variations are described and recommended depending on the sampling goal. For example, the goals described include identifying or characterizing the pattern of contamination in an aquifer, obtaining comparable period-to-period samples, and making comparisons to background when there are large seasonal fluctuations in the data.

6.2.4 Geo-statistical Studies Application

When there is spatial or temporal dependence, moving from one point to another nearby location results in values that do not change dramatically. Samples close together will tend to have more similar values than samples far apart. This is often the case in an environmental setting. Any method used to estimate an overall site mean, as well as the site variance, must properly account for the pattern of spatial continuity. Any non-random or partially random sampling scheme (including a systematic grid design) will tend to produce biased estimates if not adjusted for the degree of spatial correlation. There exist techniques to minimize the biasing impact of spatial correlation while generating reasonable estimates of the mean.

EPA has produced guidance for geostatistical soil sampling (EPA, 1996). Sampling in support of geostatistical analysis is an important topic and discussed in detail in this EPA document. One important construct in geostatistics is the variogram. The variogram is a plot of the variance of paired sample measurements as a function of the distance between samples. Samples taken on a regular grid are required to estimate the variogram. While all regular grids

tend to work reasonably well in geostatistical applications, there are differences in efficiency depending on the type of grid pattern chosen. The most common grid types include square, triangular, and hexagonal patterns. Entz and Chang (1991) evaluated 16 soil sampling schemes to determine their impact on directional sample variograms and kriging. They concluded that for their case study, grid sampling required more samples than stratified random sampling and the stratified-grid design, but the accuracy of the kriged estimates was comparable for all sampling designs. They also found that the variograms that were estimated from sample data collected from stratified and grid designs led to the same conclusion about the spatial variability of the soil bulk density (the subject of the study).

6.2.5 Hot Spot Problem Application

One application for using grid sampling that is widely encountered in environmental settings is in the spatial context of searching for hot spots. The problem can be formulated several ways:

- What grid spacing is needed to hit a hot spot with specified confidence?
- For a given grid spacing, what is the probability of hitting a hot spot of a specified size?
- What is the probability a hot spot exists when no hot spots were found by sampling on a grid?

For this application, sampling over a gridded area at the nodes is used to search for an object(s) of interest or, alternatively, to be able to state that an object of a specified size cannot exist if a grid node was not intersected. For example, the sampling goal may be to find if at least one 55-gallon drum is buried in an area. Optimal grid spacings for the hot spot problem have been worked out for a range of relative object sizes and orientations. The hot spot problem is discussed extensively in Gilbert (1987, pp. 119-131). Gilbert finds that the triangular or hexagonal design maximizes the spatial coverage of the grid more so than a square or rectangular design for locating hot spots.

In summary, if nothing is known about the spatial characteristics of the target population, grid sampling is efficient in finding patterns or locating rare events. If there is a known pattern or spatial or temporal characteristic of interest, grid sampling may have advantages over other sampling designs depending on what is known of the target population and what questions are being addressed by sampling.

6.3 WHAT ARE THE BENEFITS OF USING SYSTEMATIC/GRID SAMPLING?

Systematic/grid sampling has the following benefits:

- Uniform, known, complete spatial/temporal coverage of the target population is possible. A grid design provides the maximum spatial coverage of an area for a given number of samples.

- Design and implementation of grids is relatively straightforward (only a calculator and measuring device is required to implement) and has intuitive appeal; field procedures can be written simply. Once an initial point is located, the regular spacing allows field teams to easily locate the next sampling point.

- Multiple options are available for implementing a grid design. Often, sampling programs are executed in phases. The initial phase uses broad-scale grids to look for any kind of activity or hit. Once the general area or time frame of the activity of interest has been identified, smaller-scale grids are used to refine the estimates. Alternatively, during a single phase, the total area can be subdivided into areas based on the likelihood of finding properties of interest and different grid spacings used in each sub-area. In addition, one can overlay multiple grids, orient multiple grids in opposite directions, intermix fine-mesh grids with large-mesh grids, and still maintain fixed spacing required for certain applications, such as estimating the correlation function (i.e., variogram). These variations require adjustments to the standard formulas for estimating sample size and population parameters.

- Regularly spaced or regularly timed samples allow for spatial and temporal correlations to be calculated. If independent populations cannot be assumed and there are distinct features in the population to be sampled, regular spacing of samples is the only option for estimating the features and making predictions of unsampled areas.

- Some regulations require samples that are collected using a grid pattern.

- Grid designs can be implemented with little to no prior information about a site. The only inputs required are the total area to be covered and the number of samples (or alternatively, the grid spacing) to be used. Grid sampling is often used for pilot studies, scoping studies, and exploratory studies.

Many studies have been performed using simulated data sets to compare the efficiency of alternative sampling designs. All such studies conclude that the overall performance of the design is influenced as much by particular features in the population to be sampled along with the estimators used for estimating population parameters of interest as the type of design chosen.

Still, it is informative to observe the comparative advantages and disadvantages found by the various researchers for the various designs.

In a study on trace elements in contaminated soil to assess the impact of contaminated soil on the environment and on agricultural activities, Wang and Qi (1998) found that given a certain sampling density, systematic sampling had better estimation performance than either a cellular stratified or a random sampling design.

In a study on assessing the percent cover of crop residue to estimate soil erosion, Li and Chaplin (1995) found that systematic sampling was more precise than random sampling for both corn and soybean residue in most cases. Crop residue is plant material left on the field surface after harvest. Measuring the crop residue cover on the soil surface is essential in the management of soils to reduce erosion. Li and Chaplin laid grid frames on top of a picture taken of fields with corn and soybean residue. The image was then read into a computer program that randomly changed the position of the grid on the picture. Light densities recorded the reading of coverage at each node. The grid design compared favorably to a design where random locations were sampled for coverage readings, using the same number of sampling points as used in the systematic sampling.

In another study, Li and Chaplin (1998) considered both one- and two-dimensional sampling designs for estimating crop residue coverage. While widely used, no rigorous study exists on the precision of the line transect method. Li and Chaplin used a computer-generated virtual field surface and applied various sampling designs. They found the square grid was more precise than the line transect methods because of the smaller coefficient of variation over a wide range of sampling points and residue cover.

6.4 WHAT ARE THE LIMITATIONS OF USING SYSTEMATIC/GRID SAMPLING?

Limitations of systematic/grid sampling include the following:

- Systematic/grid sampling is not as efficient as other designs if prior information is available about the population. Such prior information could be used as a basis for stratification or identifying areas of higher likelihood of finding population properties of interest.

- If the population properties of interest are aligned with the grid, systematic/grid sampling raises the possibility of an overestimation or underestimation of a population characteristic. Caution should be used if there is a possibility of a cyclical pattern in the unit or process to be sampled that might match the sampling frequency. For example, one would not want to take air samples every Monday morning if a nearby plant always pressure-cleaned the duct work on Monday morning.

- As mentioned earlier, a single systematic sample cannot be used to get a completely valid estimate of the standard error of the mean, i.e., Var( ), without some assumptions about the population. Several approximate methods have been proposed by Wolter (1984) and illustrated by Gilbert (1987, pp. 100-102). One option is to take multiple sets of systematic samples, each with a randomly determined starting point, and calculate an empirical estimate of the standard error of the mean. Methods for estimating Var( ) developed for random sampling plans can be used with confidence only when the population is in random order.

6.5 HOW DO YOU IMPLEMENT SYSTEMATIC/GRID SAMPLING?

Systematic sampling designs are relatively straightforward to implement. You need to know how many samples to take and where to take them.

6.5.1 How do you decide how many samples to take?

Many of the sample size formulas provided for simple random sampling (i.e., the sample size formula for estimating a mean provided in Chapter 4) can be used for systematic sampling as long as there are no strong cyclical patterns, periodicities, or significant spatial correlations between pairs of sample locations. For the hot spot problem, there are nomographs (Gilbert, 1987 pp. 122-126) and a computer program called ELIPGRID PC (Davidson, 1995) for calculating the optimal grid spacing for a hot spot of prespecified size and shape with a specified confidence of finding the hot spot. Li and Chaplin (1998, pp. 650-651) discuss how to design grid sampling patterns with the least number of sampling points to achieve required precision based on results.

6.5.2 How do you decide where to take samples?

There are many variations on patterns for regular spacing of systematic samples. Patterns include square, rectangles, triangles, circles, and hexagons. Basic geometry can be used to determine internodal spacing. For example, for the two-dimensional sampling problem, EPA has detailed guidance on how to locate samples using a systematic sampling design (EPA, 1989, pp. 5-5–5-12). Figure 19, taken from that document, summarizes how to lay out a square grid. Once a sample size *n* and the area *A* to be sampled have been specified, equations 6-1 and 6-2 can be used to calculate the spacing between adjacent sampling locations. For the square grid, the distance L between the vertical and horizontal parallel lines is

$$L = SQRT (A/n) \qquad (6\text{-}1)$$

For the triangular grid, the distance L becomes

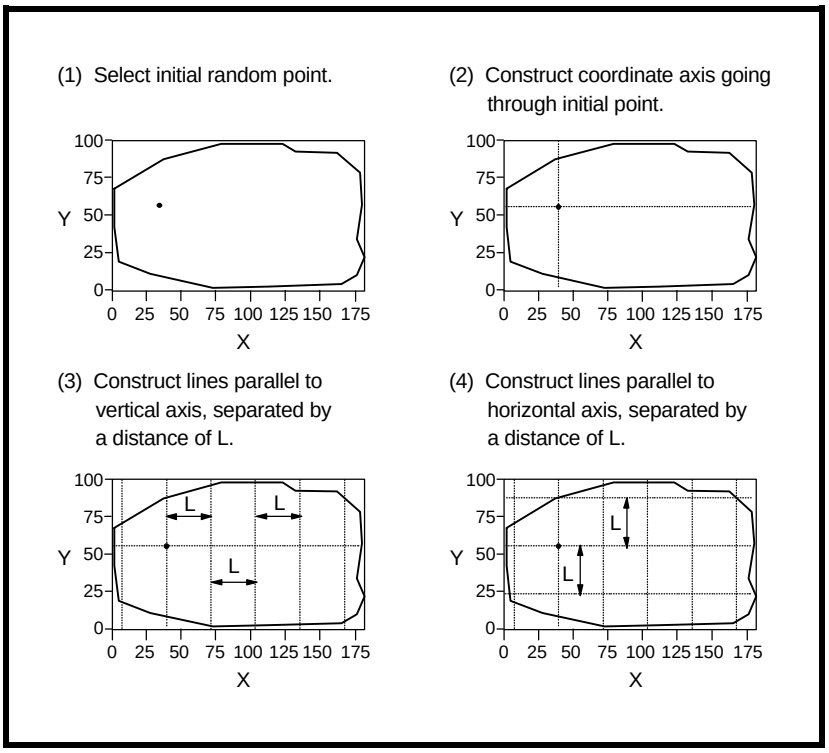


Figure 19. Locating a Square Grid Systematic Sample (EPA, 1989, pp. 5-8)

$$L = SQRT(A/(.866 * n)) \qquad (6\text{-}2)$$

For one-dimensional sampling, the procedure is even simpler. An example of a one-dimensional sampling problem is shown in Figure 20.

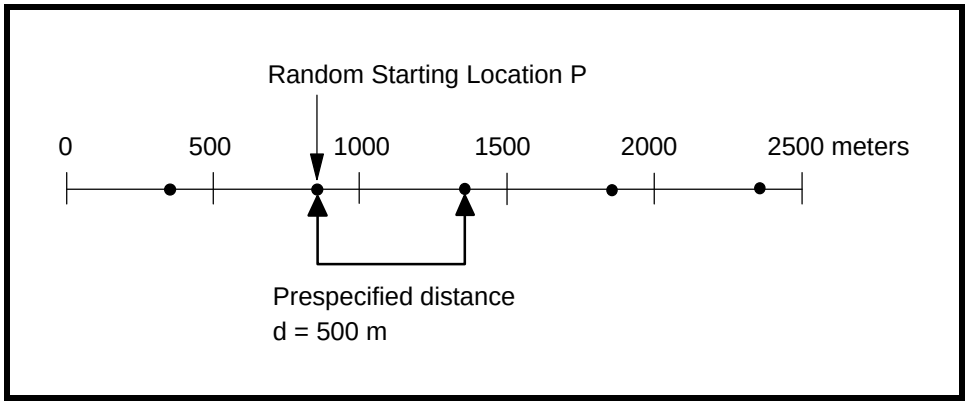


Figure 20. Choosing Sampling Locations for a Systematic Sample Along a Transect (Gilbert, 1987, p. 92)

The complexities for the one-dimensional problem come in the application. For example, the line transect method is used extensively by USDA technicians as a quick means to estimate agricultural conditions, such as plant coverage. To conduct a measurement in a certain area, a cord with 50 to 100 equally spaced beads is stretched diagonally across the crop rows. Using the same point on each bead—for example, the leading edge—those beads are counted that have the plant characteristic or interest under them when viewed directly from above. This count is divided by the total number of beads on the cord to give an observation of the percent occurrence. An average of three to five observations in the area is used to estimate field totals. The transect length, size of the cord, and marker spacing are part of the protocol.

For more discussion of the diagonal line transect method, refer to the MidWest Plan Service (MWPS, 1992). Also, see Li and Chapin (1998) for more detailed information on implementing this method.

6.6 UNDER WHAT CONDITIONS ARE OTHER DESIGNS COMMONLY USED IN CONJUNCTION WITH SYSTEMATIC/GRID SAMPLING?

Systematic sampling can be used in place of random sampling in many of the designs discussed in this document. For example, within a stratum, a grid pattern can be laid out. A grid pattern can be used in Rank Set Sampling to select initial units to rank. The key criteria for using a systematic design in lieu of a random design is that a random starting location must be identified for the selection of the initial unit and the grid layout cannot coincide with a characteristic of interest in the population.

The Environmental Monitoring and Assessment Program discussed earlier uses a sampling strategy that has multiple stages and involves aspects of stratified and systematic sampling. The first stage of the design is a triangular grid covering the conterminous United States. The grid is randomly situated over the U.S. land mass; the inter-point distance along the grid is approximately 27 km and the ratio of area to number of grid points is approximately 635 km^2 per grid point. The grid design is good for measuring those ecological resources that do not change position over the time of the survey and that need to be sampled repeatedly over time. The multistage design permits the design to be tailored to the resources of interests and purposes of the reporting. During the first stage, data may be collected at random sample grid points; on the basis of these data, informed choices can be made for the definition, stratification, and so on of second and lower stage units. In preparation for the second stage, a randomly placed hexagonal template is constructed over the region. The typical size of the template is 16 hexagons per grid point (Cox et al., 1995, p. 13).

The combination of systematic and random sampling was demonstrated in a study by Cailas et al. (1995) in proposing a methodology for an accurate estimation of the total amount of materials recycled. One objective of this comprehensive study of the recycling infrastructure in Illinois was to make an accurate estimation of the amount of total material recycled. It was found that responses from a small number of critical facilities were essential for an accurate estimation of

the total amount of material recycled. The combined design consisted of systematically sampling the critical facilities and randomly sampling the remaining ones. This application yielded an accurate estimate with less than 1 percent difference from the actual amount recycled. This was done with only 15 percent of the total number of recycling facilities included in the critical facilities sub-population.

In some cases, designs are distinguished more by semantics rather than real structural or procedural differences. For example, systematic sampling can be considered a form of stratified random sampling in which the boundaries of the strata are arbitrarily defined (rather than using prior information) and only one random sample is taken per stratum (Gilbert,1987).

6.7 EXAMPLE

This example is taken from EPA (1989, 1992). There were thirty samples taken from an area of 14,025 m^2. This area is shown in gray in Figure 21.

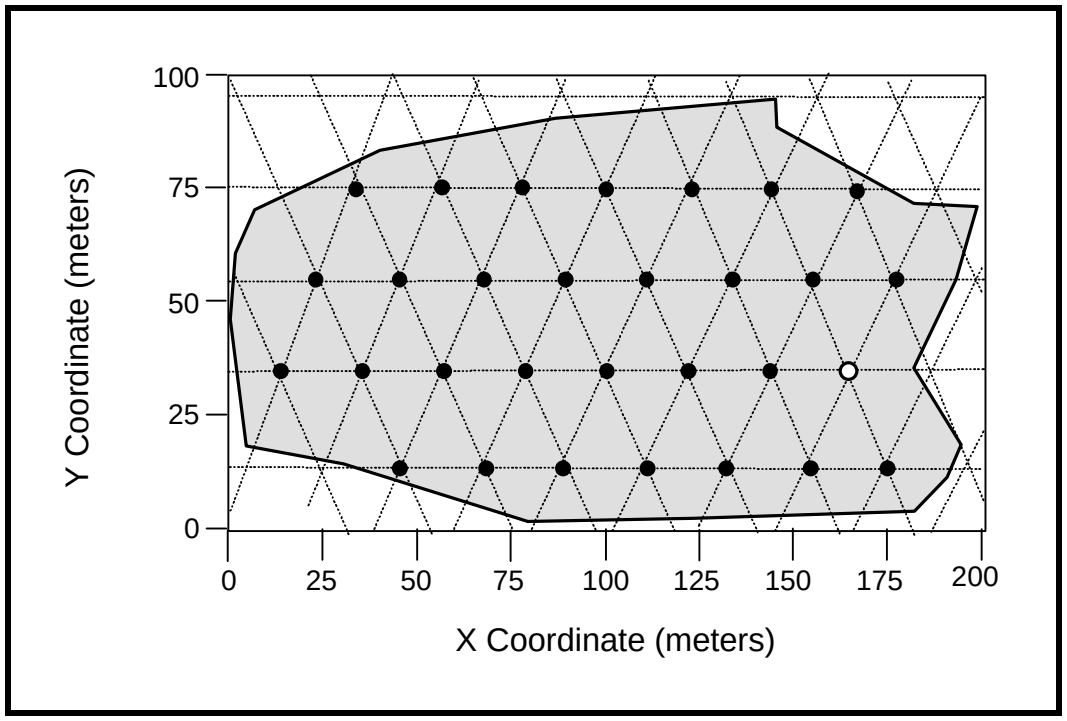


Figure 21. Map of an Area to Be Sampled Using a Triangular Sampling Grid (EPA, 1992, p. 5.4)

The following steps are performed:

1. The boundaries for the problem are determined to be $X_{min} = 0$, $Y_{min} = 0$, $X_{max} = 200$, and $Y_{max} = 100$.

2. A random number generator is used to obtain two random numbers (RND_1 and RND_2) between 0 and 1. For purposes of this example, the two numbers drawn were 0.820 and 0.360.
3. The random start location (X, Y) is obtained by using the formulas

$$X = X_{min} + RND_1 * (X_{max} - X_{min})$$
$$Y = Y_{min} + RND_2 * (Y_{max} - Y_{min})$$

4. Substituting the values from Steps 1 and 2 results in the location (X=164, Y=36). This point is shown as an open circle in Figure 21.
5. Using the formula for *L* we get

$$L = SQRT(Area/.866*n)$$

$$L = SQRT(14{,}025/0.866*30) = 23.23 = 23$$

6. A line parallel to the X axis through the point (164,36) is drawn; points are marked off 23 m apart from this line as shown in Figure 21.
7. The midpoint between the last two points along the line is found and a point is marked at a distance 0.866*23 = 19.92 = 20 m perpendicular to the line at that midpoint. This point is the first sample location on the next line.
8. Points at distance L=23 m apart are marked on this new line.
9. Steps 6 and 7 are repeated until the triangular grid is determined.

There are now exactly 30 locations marked off in a triangular pattern.

CHAPTER 7

RANKED SET SAMPLING

7.1 OVERVIEW

This chapter describes and illustrates ranked set sampling, an innovative sampling design originally developed by McIntyre (1952). The unique feature of ranked set sampling is that it combines simple random sampling with the field investigator's professional knowledge and judgment to pick places to collect samples. Alternatively, field screening measurements can replace professional judgment when appropriate. The use of ranked set sampling increases the chance that the collected samples will yield representative measurements; that is, measurements that span the range of low, medium, and high values in the population. This results in better estimates of the mean as well as improved performance of many statistical procedures such as testing for compliance with a risk-based or background-based (reference-based) standards. Moreover, ranked set sampling can be more cost-efficient than simple random sampling because fewer samples need to be collected and measured.

The use of professional judgment in the process of selecting sampling locations is a powerful incentive to use ranked set sampling. Professional judgment is typically applied by visually assessing some characteristic or feature of various potential sampling locations in the field, where the characteristic or feature is a good indicator of the relative amount of the variable or contaminant of interest that is present. For example, the relative amounts of a pollutant in randomly selected sampling spots may be assessed based on the degree of surface or subsurface soil staining, discoloration of soil, or the amount of plant defoliation in each spot. Similarly, the yield of a plant species in each of randomly selected potential 1 meter x 1 meter field plots may be visually assessed based on the density, height, or coloration of vegetation in each plot. This assessment ranks the visually assessed locations from smallest to largest with respect to the variable of interest; it is then used as described in this chapter to determine which spots to actually sample.

In some situations, a more accurate assessment of the relative amounts of a pollutant present at field locations can be provided by an inexpensive field screening measurement. Indeed, the sensitivity and accuracy of *in-situ* detectors has increased greatly in recent years Some examples include the following:

- Using ultraviolet fluorescence in the field to measure (screen) for BTEX (benzene, toluene, ethyl benzene, and xylene) and PAHs (polyaromatic hydrocarbons) in soil.
- Using X-ray fluorescence in the field to measure lead or metals in soil.
- Using reflectance intensity of near-infrared electromagnetic radiation as recorded in a remotely sensed digital image to approximate vegetation density on the ground.
- Using total organic halide (TOX) measurements of soil as a screening measurement for volatile organic solvents.

- Using remotely sensed information (aerial photographs and/or spatially referenced databases as found in a GIS) to identify locations to be studied.
- Using distance along a pipeline (longer distance implying lower levels of a contaminant) to approximate the relative concentrations of a contaminant at various distances

A simple ecological example will illustrate the ranked set sampling approach (after Stokes and Sager, 1988). A more detailed lead contamination example follows in Section 7.2. The recommended step-by-step process for setting up an ranked set sampling design is presented in the appendix to this chapter. Suppose the average volume of trees on a property needs to be estimated. Begin by randomly selecting two trees and judge by eye which tree has the most volume. Mark the *smaller* tree to be carefully measured for volume and ignore the other tree. Next, randomly select another two trees. Mark the *larger* of these two trees and ignore the other tree. Then repeat this procedure, alternatively marking the smaller of the first two trees, then the larger of the second two trees. Repeat this procedure a total of 10 cycles for a total of 40 trees. Twenty of the trees will have been marked and 20 ignored. Of the 20 marked trees, 10 are from a stratum of generally smaller trees and 10 are from a stratum of generally larger trees. Determine the volume of each of the 20 marked trees by careful measurement and use that measurement to estimate the average volume per tree on the lot. In this illustration there were 10 cycles and 2 trees marked per cycle. In practice, the number of trees marked per cycle (the "set size") and the number of cycles is determined using a systematic planning process, as illustrated in the Appendix.

7.1.1 Example of Using Ranked Set Sampling to Estimate The Mean Lead Concentration in Soil

Suppose a future residential area is suspected of having lead concentrations in surface soil that exceed background concentrations. As part of the risk assessment process, the soil of the area must be sampled to estimate the mean lead concentration. Prior studies have shown that X-ray Fluorescence (XRF) measurements of lead in soil obtained using a hand-held *in-situ* detector closely correlate with laboratory measurements of lead in soil at the same locations. Furthermore, it was determined that the cost of taking the XRF measurements was very low compared to the cost of laboratory measurements for lead. Hence, it was decided to use ranked set sampling instead of simple random sampling (see the Appendix for guidance on how to determine if ranked set sampling is preferred over simple random sampling).

Suppose the systematic planning process employed determined that $n = 12$ soil samples should be collected in order to have 95 percent confidence that the estimated mean would be within 25 percent of the true mean. Also, to properly compute the variance of the estimated mean, $m = 3$ samples (the "set size") were collected in each of $r = 4$ cycles to obtain the required $n = m \times r = 3 \times 4 = 12$ samples. A method to determine m and r is provided in the Appendix.

The ranked set sampling method for determining the three field locations to be sampled is as follows:

1. Use simple random sampling to randomly select $m^2 = 3^2 = 9$ locations on the property. Randomly divide the 9 locations into m sets of size m (3 sets of size 3). In Figure 22 the first set of three locations is denoted by “Set 1," the second set by “Set 2," and the third set by “Set 3.”

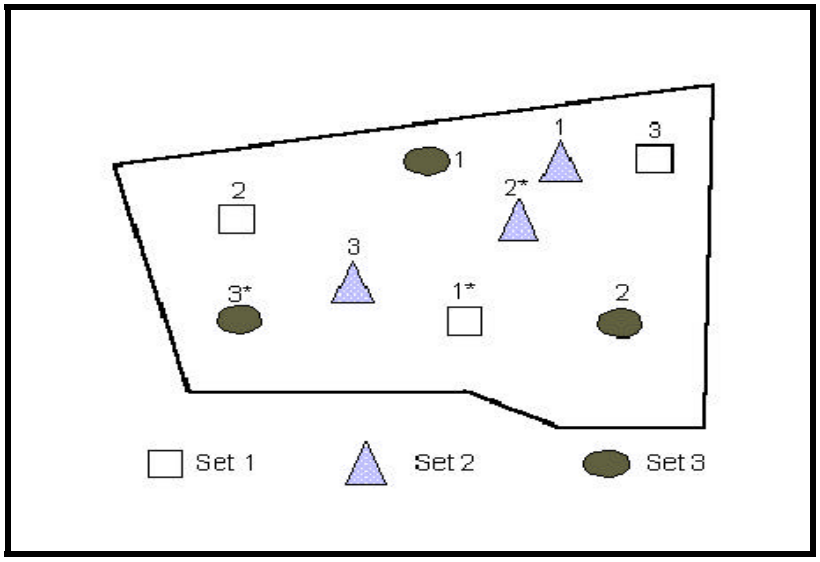


Figure 22. Using Ranked Set Ssampling to Select Three Locations - those identified by * - Where a Soil Sample Will Be Collected for Measurement of Lead in the Laboratory

2. Consider the three locations in Set 1. Make an XRF measurement at each of the three locations and label the locations *1*, *2*, and *3* to indicate the smallest, middle, and largest XRF measurement, respectively. Collect the first soil sample at label *1* in Set 1: this location has the smallest XRF lead measurement in Set 1 (labeled *1** in Figure 22).

3. Consider the three locations in Set 2 and make an XRF measurement at each of those locations. Collect the second soil sample at label *2* in Set 2: this location has the second highest XRF measurement in Set 2 (labeled *2** in Figure 22).

4. Consider the three locations in Set 3 and make an XRF measurement at each of the three locations in that set. Collect the third soil sample at label *3* in Set 3: this location has the highest XRF measurement in Set 3 (labeled 3* in Figure 22).

This procedure is repeated $r = 4$ times to obtain the $n = m \times r = 3 \times 4 = 12$ soil samples needed. This replication process is needed to estimate the variance of the estimated mean (see the Appendix for the computational formula). In practice, if professional judgment is used to rank the locations in each set, the set size ($m = 3$ in this example) is usually between 2 and 5. Larger values of m make it more difficult to accurately rank the locations within each set. However, *m*

values larger than 5 may be practical if field locations are ranked using screening measurements. In general, larger set sizes are desirable because they result in more precise estimates of the mean.

Note that the above example is a balanced ranked set sampling design (i.e., the same number of field locations are sampled for each rank). Unbalanced ranked set sampling designs can also be used, as discussed in Section 7.5.2 and the Appendix to this chapter.

7.2 UNDER WHAT CONDITIONS IS RANKED SET SAMPLING APPROPRIATE?

Ranked set sampling is appropriate when the following conditions hold:

- The cost of laboratory measurements is high relative to the cost of using screening measurements or professional judgment in the field to determine the relative magnitudes of contamination in randomly selected field plots.

- Professional judgment or field screening measurements can accurately determine the relative magnitudes of contamination among randomly selected field locations.

- A more precise estimate of the mean or a more powerful test for compliance is needed than can be achieved for a fixed budget if simple random sampling were used in place of ranked set sampling.

A process whereby costs and accuracy of ranking field locations is considered in setting up an ranked set sampling design is provided in the appendix to this chapter.

7.3 WHAT ARE THE BENEFITS OF USING RANKED SET SAMPLING?

A major benefit of ranked set sampling is that it will yield a more precise estimate of the mean than if the same number of measurements is obtained using simple random sampling (McIntyre, 1952; Gilbert, 1995; U.S. EPA, 1995; Johnson et. al., 1996; Muttlak, 1996). Table 18 illustrates this for the normal distribution. For example, suppose the distribution of the variable of interest is normal with a true mean of 1 and a coefficient of variation (CV = standard deviation divided by the mean) of 0.50. Furthermore, suppose our goal is to obtain enough laboratory measurements to be 95 percent confidence that the estimated mean is within 25 percent of the true mean. Table 18 indicates that simple random sampling will require 16 samples, but if ranked set sampling is used with a “set size” of 2, then only 12 samples are needed, reducing sampling and laboratory costs by 25 percent. If the cost of using professional judgment or field screening measurements is considerably less than the cost of laboratory measurements, then there is a strong motivation to use ranked set sampling rather than simple random sampling. Note in Table 18 that when high precision in the estimated mean is required, the number of samples needed is dramatically reduced as the set size increases.

Ranked set sampling also has several other benefits: the estimated mean of ranked set sampling data is a statistically unbiased estimator of the true mean; ranked set sampling provides increased ability to detect differences in means or medians of two populations (e.g., site and background populations); ranked set sampling can be used in place of simple random sampling in other sampling designs such as stratified random sampling and composite sampling; and ranked set sampling can be used to obtain more representative data for purposes other than estimating a mean. Such purposes include computing a confidence limit on the median of a population (Hettmansperger, 1995), testing for differences in the medians of two populations (Bohn and Wolfe, 1992, 1994), conducting simple tests to check for compliance with a fixed remediation concentration limit (Hettmansperger, 1995; Koti and Babu, 1996; Barabesi, 1998), estimating the slope and intercept of a straight line relationship (Muttlak, 1995), estimating the ratio of two variables (Samawi and Muttlak, 1996), and estimating the means of several populations in an experimental setting (Muttlak, 1996).

Table 18. Comparing the Number of Samples Required If Ranked Set Sampling Instead of Simple Random Sampling Is Used to Estimate the Mean with Specified Precision and 95% Confidence*

		Required Precision of the Estimated Mean with 95% Confidence			
CV**	**RSS Set Size (m)**	**10%**	**15%**	**25%**	**50%**
0.50	SRS	97	43	16	4
	RSS - 2	66	30	12	4
	RSS - 3	51	24	9	3
	RSS - 5	35	20	10	5
0.707	SRS	193	86	31	8
	RSS - 2	32	60	22	6
	RSS - 3	102	45	18	6
	RSS - 5	70	35	15	5
1.0	SRS	385	171	62	16
	RSS - 2	262	118	42	12
	RSS - 3	201	90	33	9
	RSS - 5	140	65	25	10

* Adapted from Table 1 in Mode et al. (1999). Tabled values were derived assuming there are no errors in ranking field locations. ** CV = Coefficient of Variation = standard deviation of the data divided by the mean.

7.4 WHAT ARE THE LIMITATIONS OF USING RANKED SET SAMPLING?

When the objective of sampling is to estimate the mean, ranked set sampling has very few limitations. Ranked set sampling should therefore always be considered for estimating a mean instead of simple random sampling because ranked set sampling is always more cost efficient than simple random sampling (Patil et al., 1994, p. 175) when the cost of ranking potential sampling locations in the field is negligible or very low compared to the cost of laboratory measurements. Guidance on setting up an ranked set sampling design that includes cost considerations, including those due to ranking, is provided in the appendix to this chapter.

The precision of the mean computed using ranked set sampling data is maximized (i.e., the variance of the estimated mean is minimized) if there are no errors in ranking field locations. However, even when professional judgment or field screening methods cannot accurately rank field locations, ranked set sampling will perform on average as well as simple random sampling in estimating the mean.

In ranked set sampling, the field locations being compared (ranked) are supposed to be randomly located over the population. Field locations within a set may, however, be purposely clustered in close proximity, perhaps to decrease the effort of taking screening measurements or to increase the accuracy of using judgment to visually rank the plots. In this case, the precision of the estimated mean obtained using ranked set sampling data may be reduced somewhat, but it will never (on average) be less than the mean obtained through simple random sampling. To reduce or eliminate this decrease in precision, divide the population into portions of equal size that have no well-defined gradients and an equal number of samples are selected within each portion (McIntyre, 1952).

If ranked set sampling data are used to test hypotheses, the data computations may differ from data obtained using simple random sampling. For example, suppose differences in the medians of two populations based on ranked set sampling data from both populations are being tested using the Wilocoxon Rank Sum test. The data computations to do the Wilocoxon Rank Sum test as described in Bohn and Wolfe (1992, 1994) should be used rather than the usual computational methods that are suitable to simple random sampling. If ranked set sampling data will be used to conduct tests of hypotheses or to compute confidence intervals on means or other statistical parameters, guidance from a statistician familiar with ranked set sampling should be sought.

7.5 HOW DO YOU IMPLEMENT RANKED SET SAMPLING?

7.5.1 How do you decide what sample size to use to estimate the mean?

Most methods in the statistical literature for determining sample size for estimating the mean were developed assuming that sampling locations are identified using simple random sampling rather than ranked set sampling. In general, ranked set sampling requires fewer samples

than simple random sampling because ranked set sampling yields more information per set of measurements. This concept was illustrated in Table 18 for the normal distribution. The appendix to this chapter provides a step-by-step process for determining the ranked set sampling sample size for estimating a mean.

Methods for computing the ranked set sampling sample size for other sampling objectives, such as testing hypotheses, are less well-developed and not yet available in the statistical literature. However, since ranked set sampling usually increases the performance of statistical procedures relative to what would be achieved if simple random sampling were used, the recommended approach is to use the same number of samples that would be needed if simple random sampling were used.

7.5.2 How do you decide where to take samples?

Locations for sample collection are determined by the ranking process using professional judgment or field screening measurements. The use of ranked set sampling to determine the field locations is illustrated in the appendix for a balanced ranked set sampling design. In a balanced ranked set sampling design, the same number of locations are collected for each rank. For example, the simple ranked set sampling lead example given in Section 7.1 was a balanced design because the design required an equal number of locations expected to have relatively low, medium, or high lead concentrations. A balanced ranked set sampling design should be used if the underlying distribution of the population is symmetric.

In an unbalanced ranked set sampling design, different numbers of locations expected to have relatively low, medium, or high concentrations are sampled. Environmental data are often asymmetric and skewed to the right; that is, with a few measurements that are substantially larger than the others. If the goal is to estimate the mean, McIntyre (1952) indicates the mean would be more precisely estimated if more locations expected to have relatively high concentrations were selected than locations expected to have relatively low or medium concentrations. This idea is discussed further by Patil et al. (1994, pp. 187-188). To illustrate an unbalanced ranked set sampling design, one could modify the lead example in Section 7.1 to collect a soil sample at twice as many locations expected to have relatively high lead concentrations as at locations expected to have relatively low or medium concentrations. When an unbalanced ranked set sampling design is used, the true mean of the population is estimated by computing a weighted mean, as described in the appendix, rather than the usual unweighted mean.

An appropriate unbalanced ranked set sampling design should increase the precision of the estimated mean of an asymmetric distribution. However, an inappropriate unbalanced ranked set sampling design for an asymmetric distribution can provide a less precise estimate of the mean than a balanced ranked set sampling design or a simple random sampling design. Kaul et al. (1995) established a method for developing an appropriate unbalanced ranked set sampling design asymmetric distributions that are skewed to the right. This method is provided in the appendix.

7.6 EXAMPLES

7.6.1 Estimating Mean Plutonium Concentrations in Soil

Gilbert (1995) illustrates the use of ranked set sampling to obtain samples for estimating the mean plutonium (Pu) concentration in surface soil at some weapons testing areas on the Nevada Test Site. Pu concentrations in soil samples are typically measured in the laboratory, and measurement is quite expensive. However, at the weapons testing areas in Nevada**,** inexpensive field measurements of Americium-241 (denoted by ^{241}Am) in surface soil can be obtained using an *in-situ* detector called the FIDLER (Field Instrument for the Detection of Low Energy Radiation)**.** Past studies had shown that in areas of high soil Pu concentrations, there is a reasonably high correlation (about 0.7) between a FIDLER reading at a field location and a Pu measurement made on a 10-g aliquot for a surface (0-5 cm) soil sample collected at that spot. Moreover, the cost of a Pu measurement in the laboratory is at least 10 times greater than the cost of obtaining a FIDLER reading. Hence, using Table 20 in the Appendix to this chapter, it appears that using ranked set sampling instead of simple random sampling to determine soil locations to sample should provide a more precisely estimated mean. Gilbert (1995) illustrates how to compute the mean and its variance using data from a balanced ranked set sampling design. It should be noted that, because the distribution of Pu measurements at the study areas is typically skewed to the right, an unbalanced ranked set sampling design might produce a more precise estimated mean than a balanced ranked set sampling design.

7.6.2 Estimating Mean Reid Vapor Pressure

Nussbaum and Sinha (1997) discuss a situation where ranked set sampling appears to have great potential for cost savings. Air pollution in large cities is currently being reduced through the use of reformulated gasoline. Reformulated gasoline was introduced because of regulations that limit the volatility of gasoline, as commonly measured by the Reid Vapor Pressure (RVP). Typically, RVP is measured on samples from gasoline stations obtained using simple random sampling. RVP can be measured in the laboratory or at the pump itself. Although laboratory measurement costs are not unduly expensive, it is expensive to ship samples to the laboratory. Hence, reducing the number of samples analyzed in the laboratory could result in a large costs savings without sacrificing the assessment of compliance with the volatility regulations.

One possible way to reduce the number of samples analyzed in the laboratory is to use ranked set sampling. Measurements of RVP taken at the pump might be used to rank samples using the ranked set sampling procedure to determine which samples should be taken to the laboratory for measurement. Suppose that (1) the correlation between field RVP and laboratory RVP measurements is sufficiently high so that the ranking was very accurate and that (2) it is several times more costly to transport and measure samples in the laboratory than it is to rank samples at the pump. In this case, the number of samples measured in the laboratory could be reduced by perhaps a factor of 2 or more without reducing the ability to determine when the volatility regulators are being violated. Nussbaum and Sinha (1997) present data that strongly

suggest a very strong positive linear relationship between pump and laboratory measurements of RVP. This information may be used to justify the use of field RVP measurements to accurately rank the pump samples (see Table 19 in the Appendix). Assuming no ranking errors, Table 19 shows that if the ratio of laboratory transportation and measuring costs to ranking costs (i.e., the cost of the field RVP measurement and ranking process) is greater than 6, then ranked set sampling can be expected to yield as precise an estimate of the mean RVP as what would be obtained using simple random sampling but at less cost.

7.6.3 Estimating Mean Pool Area in a Stream

Mode et al. (1999) provided this example of a USDA Forest Service data collection effort on Pacific Northwest streams as part of a large scale monitoring project. There was interest in assessing salmon production in streams. The size of salmon habitat, particularly pool area in streams, has been linked to salmon production. Obtaining pool area by accurately and precisely measuring length and width of stream pools is time consuming and labor intensive. However, visual estimates of pool area can be obtained at much less cost. Mode et al. (1999) found that ranked set sampling estimates of the mean pool area for 20 of 21 streams were more precise than estimates of the pool area that would be obtained by physically measuring pool areas selected using simple random sampling. They also found that for over 75 percent of the streams, it would be less costly to use ranked set sampling than simple random sampling to obtain the same precision in the estimated mean pool area when pool measuring costs were at least 11 times greater than the costs of visually assessing pool area.

APPENDIX 7-A
USING RANKED SET SAMPLING

INTRODUCTION

This Appendix provides guidance on how to develop a balanced or unbalanced ranked set sampling design and how to estimate the mean and the standard deviation of the mean based on the data obtained. Developing a ranked set sampling design for the purpose of estimating the mean of the population is a two step process:

Step 1. Determine if ranked set sampling is cost effective compared to simple random sampling. This step is accomplished by considering the costs and performance of professional judgment and inexpensive field screening methods for ranking field locations.

Step 2. If ranked set sampling is expected to be more cost effective than simple random sampling, then determine the number of samples required to estimate the mean with required accuracy and confidence.

Details of how to implement Steps 1 and 2 are provided in this appendix of this chapter along with the methods for computing the mean and its standard deviation.

HOW DO YOU DECIDE IF RANKED SET SAMPLING IS MORE COST EFFECTIVE THAN SIMPLE RANDOM SAMPLING FOR ESTIMATING THE MEAN?

This section provides guidance on how to determine if ranked set sampling will be more cost effective than simple random sampling when the objective of sampling is to estimate the mean with a specified precision. It is known (Patil et al., 1994, p. 175) that ranked set sampling is more cost effective than simple random sampling for estimating the mean if the cost of using professional judgment or field screening measurements to rank potential sampling locations is negligible. This conclusion stems from the fact that fewer ranked set sampling samples than simple random sampling samples are needed to estimate the mean with specified precision. Hence, laboratory measurement costs will be lower. However, ranking potential sampling locations in the field may be costly due to factors such as spending more hours in the field, locating and training an expert to subjectively rank field locations, and purchasing and using field screening detectors. The basic question is whether the increased precision in the mean that can be obtained using ranked set sampling will compensate for the extra work and cost of ranking.

We can approximate the effect of costs on the decision of whether to use ranked set sampling or simple random sampling using Table 19. This table shows the approximate cost ratio (cost of a laboratory measurement divided by the cost of ranking a field location) that must be exceeded before ranked set sampling will be more cost effective than simple random sampling to estimate the mean with a desired level of precision. The cost ratio that must be exceeded depends

on the set size, *m* (number of locations sampled in each of the *r* ranked set sampling cycles) and on the distribution of the population of laboratory measurements. Table 19 gives approximate cost ratios for normally and exponentially distributed measurements when there is either no ranking error or a substantial likelihood that ranking errors will occur. Table 19 shows that for a given set size, the cost ratios that apply when there is substantial ranking error are almost double the ratios when there is no ranking error. Also, the cost ratios are approximately 50% higher for the exponential distribution than for the normal distribution in most cases. This situation indicates that a balanced ranked set sampling design loses some of its advantage over simple random sampling when the distribution of the measurements is asymmetrical and highly skewed to the right so that most measurements are relatively low in value but a few are much larger.

Suppose that practical aspects of ranking in the field lead to using a relatively small set size of $m = 3$ and that prior studies at the site of interest indicate that laboratory measurements for the contaminant of interest are likely to be approximately normally distributed. Since the normal distribution is symmetric, a balanced ranked set sampling design will be used (a balanced design is defined in Section 7.5.2). If we expect no errors in ranking field locations, the ratio of laboratory measuring costs (per sample) to ranking cost (per field location) must be greater than approximately 3 in order for ranked set sampling to be more cost effective than simple random sampling; that is, for the total cost of ranked set sampling to be less than the total cost of simple random sampling to estimate the mean with a desired specified precision. If there is substantial ranking error and $m = 3$ is used, the cost ratio must be greater than 6 for ranked set sampling to be more cost effective than simple random sampling. However, if past studies indicate that the measurements are more likely to have a distribution that is skewed to the right, the cost ratios for the exponential distribution in Table 19 should be used. If (1) field concentrations are exponentially distributed, (2) a balanced ranked set sampling design with set size $m = 3$ is used, and (3) no ranking errors are expected, then the cost ratio must be greater than 6 before ranked set sampling is more cost effective than simple random sampling for estimating the mean. If substantial ranking errors are expected, then the cost ratio must be greater than 9.

Table 19. The Approximate Cost Ratio* for Estimating the Mean

Data Distribution	Degree of Ranking Error	Set Size m = 2	Set Size m = 3	Set Size m = 5
Normal	None	4	3	3
Normal	Substantial**	7	6	6
Exponential	None	6	5	4
Exponential	Substantial**	10	9	9

Constructed from Figure 3 in Mode et al. (1999).

*Cost of a laboratory measurement divided by the cost of ranking a field location.

**Quantitatively, a "substantial degree of ranking error" is defined here as ranking error modeled as a normal distribution that has mean zero and variance 0.5, as used in Figure 3 of Mode et al. 1999.

Although cost ratios of the type developed by Mode et al. (1999) are not available for distributions other than the normal and exponential, the ratios in Table 19 provide general guidance for symmetric and right-skewed distributions. Note that the cost ratios in Table 19 were developed assuming that a balanced ranked set sampling design will be used. If the distribution of laboratory measurements is expected to be skewed to the right, then an unbalanced ranked set sampling design will be more efficient than a balanced ranked set sampling design. That is, the cost ratios for an unbalanced ranked set sampling design would be smaller than those given in Table 19 for the exponential distribution for a balanced ranked set sampling design.

The cost ratios in Table 19 can be used when field locations are ranked using either professional judgment or field screening measurements. Table 20 provides cost ratios (from Mode et al., 1999, Figure 4) for balanced ranked set sampling designs with set sizes *m* equal to 2, 4, 6, and 8 that are applicable when there is quantitative information on the correlation between the field screening measurement at a location and the measurement obtained in the laboratory for a sample collected at the field location. If the field screening measurement is a good predictor of the corresponding laboratory measurement, then the correlation between the two measurements will be close to 1 and no or very few ranking errors will occur. A correlation of exactly 1 implies no ranking errors. If the screening measurement has absolutely no ability to predict the value of the laboratory measurement, then the correlation will be zero.

Table 20. Approximate Cost Ratio* for Estimating the Mean when Field Screening Measurements Are Used to Rank Field Locations**

Correlation (Degree of Ranking Error)	Set Size m = 2	Set Size m = 4	Set Size m = 6	Set Size m = 8
1.0 (No ranking error)	5	3	2	2
0.9	6	5	5	5
0.8	7	8	8	9
0.7	12	12	14	16

*Cost of a laboratory measurement divided by the cost of ranking a field location.
**Cost ratios are from Figure 4 of Mode et al. (1999) and were derived assuming the field screening measurements and the measurements in the laboratory have a bi-variate normal distribution.

If the correlation between the screening and laboratory measurements is close to 1, then the information gained by ranked set sampling via the ranking process increases appreciably compared to simple random sampling. Hence, the cost ratio need not be so large for ranked set sampling to be worth the extra effort and cost of ranking. For example if the correlation is 1, indicating no ranking errors, then the cost ratio can be as small as 2 or 3 for set sizes of $m = 4$ or larger. But ranking errors will occur if the correlation is 0.8 or smaller, and the additional information obtained using ranked set sampling will be reduced compared to simple random

sampling. Consequently, the cost ratio that must be exceeded for ranked set sampling to be more cost effective than simple random sampling is relatively high (8 or more).

Tables 19 and 20 permit summary statements like the following (adapted from Mode et al., 1999, pp.188-189): If the cost for a laboratory measurement is about six times that of a screening measurement or professional judgment determination, and given that past data sets have been fairly normally distributed, then ranked set sampling will be more cost effective than simple random sampling unless the chosen ranking method will result in substantial ranking errors (Table 19) or is based on a field screening measurement that is not very highly correlated (Table 20).

HOW DO YOU DETERMINE THE NUMBER OF RANKED SET SAMPLING SAMPLES NEEDED TO ESTIMATE THE MEAN?

The section below provides definitions and information about the relative precision of ranked set sampling as compared to simple random sampling. Is this a title of something? This information is used in the next two sections, which give methods for approximating the number of samples (sample size) needed for balanced and unbalanced ranked set sampling designs, respectively.

What is the relative precision of ranked set sampling to simple random sampling?

The relative precision of ranked set sampling to simple random sampling is defined to be

$$RP = Var(\bar{x}_{SRS}) / Var(\bar{x}_{RSS}) \qquad 7.1A$$

where

$Var(\bar{x}_{SRS})$ = variance of the estimated mean of the laboratory measurements if simple random sampling is used to select sampling locations, and

$Var(\bar{x}_{RSS})$ = variance of the estimated mean of the laboratory measurements if ranked set sampling is used to select the sampling locations.

Note from Equation (7.1A) that values of relative precision greater than 1 imply that $Var(\bar{x}_{RSS})$ is less than $Var(\bar{x}_{SRS})$, in which case ranked set sampling should be considered for use instead of simple random sampling, assuming the applicable cost ratio in Table 19 or 20 is exceeded.

It is known (Patil et al., 1994) that the relative precision of ranked set sampling to simple random sampling is always equal to or greater than 1 when a balance design is used, regardless of the shape of the distribution of the laboratory measurement data. This means that $Var(\bar{x}_{RSS})$ is always expected to be less than $Var(\bar{x}_{SRS})$, a rather remarkable result. To be more specific, if a balanced ranked set sampling design is used, then

$$1 \leq RP \leq (m + 1) / 2 \qquad 7.2A$$

where *m* is the set size. For example, if *m* = 2, then the value of relative precision is between 1 and 1.5, and if *m* = 3, then relative precision is between 1 and 2. The particular value of relative precision for any given study population depends on the distribution of the laboratory measurements. The upper bound of the relative precision, (*m*+1)/2, is achieved when the distribution of the measurements is rectangular. The relative precision lies between 1 and (*m*+1)/2 for all other distributions. The lower bound for the relative precision, 1, occurs when ranking is completely random, that is, when professional judgment or field screening measurements have no ability whatsoever to correctly rank field locations.

The sample size procedures given below require that a value for the relative precision be specified. Patil et al. (1994, Table 1, p. 176) provide values of the relative precision for balanced ranked set sampling designs for normal, rectangular, beta, gamma, Weibull, exponential and several other distributions for set sizes m of 2, 3, 4, and 5. Patil et al. (1994, Table 2, p. 177) also provides values of the relative precision for balanced ranked set sampling designs for lognormal measurement distributions for set sizes m between 2 and 10. A portion of their relative percisions for the lognormal distribution are provided here in Table 21. It should be noted that the relative precision values in Table 21 for lognormal distributions with CV = 0.10 are

Table 21. Relative Precision (RP)* of Ranked Set Sampling to Simple Random Sampling for Lognormal Distributions

Set Size (m)	CV = 0.1**	CV = 0.3	CV = 0.5	CV = 0.8	CV = 1.3
2	1.5	1.4	1.4	1.3	1.2
3	1.9	1.8	1.7	1.5	1.3
4	2.3	2.2	2.0	1.8	1.5
5	2.7	2.6	2.3	2.0	1.6
6	3.1	2.9	2.6	2.2	1.7
7	3.6	3.3	2.8	2.4	1.8
8	3.9	3.6	3.1	2.5	1.9
9	4.3	3.9	3.3	2.7	2.0
10	4.7	4.3	3.6	2.9	2.1

* Values of relative precision are from Table 2 in Patil et al. (1994, p. 177).

** CV = Coefficient of Variation for the lognormal distribution, which is defined to be $CV = [\exp(S^2) - 1]^{1/2}$, where S^2 is the variance of the natural logarithms of the data.

also appropriate for normal distributions. This occurs because a lognormal distribution with a very small CV value has a shape very similar to a normal distribution. As the CV becomes large, the lognormal distribution has a longer and longer tail extending to high data values.

Also, the relative percisions in Table 1 of Patil et al. (1994) for gamma and Weibull (asymmetrical) distributions bracket a range of relative precision values that is similar to those for the lognormal distribution for the same set sizes, m. Since in practice we usually do not know on the basis of statistical tests (unless sample sizes are very large) whether the data are best modeled by a lognormal, gamma, Weibull, or some other right-skewed distribution, the relative precision values in Table 21 for the lognormal distribution are used here to approximate the number of samples needed to estimate the mean when a balanced ranked set sampling design is used.

Sample Size for Balanced Ranked Set Sampling Designs

The procedure for approximating the sample size needed for a balanced ranked set sampling design to estimate the population mean with specified precision and confidence is as follows:

Step 1: Use the DQO process to determine the sample size, n_o, needed to estimate the mean with required accuracy and confidence assuming that simple random sampling is used to determine the sampling locations. The method for determining n_o is provided in Chapter 4.

Step 2 : Select a value of the set size, *m*. This value is usually based on practical constraints in ranking locations in the field using professional judgment or field screening measurements. It may be difficult to use professional judgement to accurately rank by eye more than 4 or 5 locations, which implies *m* should not exceed 4 or 5. Other constraints that may affect the size of *m* are time, staff, and cost considerations.

Step 3: Use the site conceptual model in conjunction with available data or information from prior studies or from new data collected at the site (from the same population) to select a value of the relative precision. Do this by first computing the estimated coefficient of variation (CV) of data collected previously from the same or very similar site using similar collection, handling, and measurement methods. Ideally, the number of data (*N*) used to compute the CV should be at least 10 and preferably more than 30. The estimated CV is computed as follows:

Estimated CV = s / $\bar{x}$ = standard deviation / mean

where

$$\bar{x} = \sum_{i=1}^{N} x_i / N$$

$$s = [\sum_{i=1}^{N} (x_i - \bar{x})^2 / (N-1)]^{1/2}$$

x_i = the i^{th} data value

and N = the number of data values used to compute the CV.

Use Table 21 with the computed value of the CV and the selected value of the set size m to determine the approximate value of relative precision.

Step 4: Compute the number of replications (cycles), *r*, as

$r = (n_o / m) \times (1 / RP)$

Step 5: Compute the total number of samples, *n*, that should be collected to estimate the mean:

$n = r \times m$

Note from Equation 7.3A that $r = n_o / m$ and hence that $n = n_o$ if RP = 1. Values of RP equal to 1 occur if the professional judgment or field screening measurements used have no ability whatsoever to correctly rank field locations, in which case ranked set sampling has no advantage over simple random sampling. In that case, the number of samples required by ranked set sampling is the same as that required by simple random sampling. The factor 1/RP [which equals $Var(\bar{x}_{RSS}) / Var(\bar{x}_{SRS})$] in Step 4 adjusts (decreases) the value of r to account for the fact that $Var(\bar{x}_{RSS}) < Var(\bar{x}_{SRS})$ whenever RP > 1.

Also, we see from Table 21 that the relative precision is closer to 1 if the selected set size m is very small (say 2) and the CV is very large (indicating a highly skewed distribution). Hence, in this situation the number of samples needed for a balanced design to estimate the mean of a highly skewed distribution will be only slightly less than the number of samples required by simple random sampling. Accordingly, if the CV is large, consideration should be given to using an unbalanced ranked set sampling design in order to determining the number of samples using the procedure.

Example

This example expands on the lead contamination example in Section 7.2 in the main text of this chapter. Suppose the goal is to estimate the mean concentration of lead in the surface soil of a residential property and that no major spatial patterns of lead concentrations are expected at the site. This suggests that simple random sampling may be considered for determining where soil

samples should be collected for measurement of lead in the laboratory. (Stratified random sampling would have been considered if major spatial patterns existed and had been identified previously.)

However, suppose past studies had indicated that measurements of lead in soil obtained in the field using a hand-held x-ray fluorescence (XRF) in-situ detector have a correlation of approximately 0.9 with laboratory lead measurements made on soil samples collected at the measured field locations. This high correlation suggests that ranked set sampling might be used instead of simple random sampling in order to reduce the number of soil samples that would need to be measured in the laboratory. To determine if ranked set sampling would be more cost-effective than simple random sampling, the cost of a laboratory measurement for lead was divided by the cost of ranking a field location using an XRF measurement to determine a measurement-to-ranking cost ratio. Suppose this cost ratio was found to be 10, such that ranking a field location is only one-tenth as costly as a lead measurement in the laboratory. Referring to Table 20 with a correlation of 0.9, we see that the computed cost ratio (10) is greater than the tabled value of 6. Hence, it appears that ranked set sampling will indeed be more cost-effective than simple random sampling. Therefore, we proceed to determine the number of samples that should be collected using the five-step process above.

Step 1: Determine the number of field samples, n_o, to estimate the true mean assuming that simple random sampling is used to identify the locations where samples will be collected.

It was determined using the method in Table 6 in Chapter 4 that using simple random sampling to determine field sampling locations would require a total of n_o = 25 soil samples to estimate the mean lead concentration with 20 percent accuracy and 95 percent confidence.

Step 2: Select the set size, m.

The set size m was selected to be m = 5. A larger value of m was not used in order to limit time spent in the field to find and rank field locations where the XRF measurements would be taken.

Step 3: Determine the relative percision of ranked set sampling to simple random sampling.

Suppose that past studies on similar residence properties had produced 50 lead measurements of the same type as will be obtained in the present study. Displaying the data graphically using probability plots and histograms indicated that the data set was only slightly skewed to the right (to high lead concentrations). The CV of the population was estimated using the 50 measurements and was found to be 0.4. Hence, entering Table 21 with m = 5, the relative precision of a balanced ranked

set sampling design was approximated to be about 2.45 (interpolating between RPs 2.6 and 2.3 for CV = 0.3 and 0.5, respectively).

Step 4: Determine r, the number of cycles of ranked set sampling.

The number of cycles, *r*, of ranked set sampling was computed as follows:

$$r = (n_o / m) \times (1 / RP) = (25/5) \times (1/2.45) = 2.04$$

which is rounded to 3.

Step 5: Compute the total number of samples needed.
The total number of samples to be collected for the balanced ranked set sampling design is

$$n = r \times m = 3 \times 5 = 15$$

as compared to the $n_o = 25$ that would have been required if simple random sampling is used.

The balanced ranked set sampling design is implemented by first identifying $m^2 = 5^2 = 25$ field locations using simple random sampling and then randomly dividing these 25 locations into 5 sets of size 5. The XRF detector ranks the five locations within the first set of five and a soil sample is collected at the location with the lowest XRF measurement. The second set of five locations is then ranked using the XRF detector and a soil sample collected at the location with the second smallest XRF measurement in that set, and so on through the five sets of five locations to obtain five soil samples. Then that process is repeated $r = 3$ times to obtain a total of $r \times m = 3 \times 5 = 15$ soil samples that are measured for lead in the laboratory. (Note that this process required a total of $m^2r = 25 \times 3 = 75$ field locations to be measured by the XRF detector.) The (unweighted) arithmetic average of the 25 lead measurements is then computed to estimate the true mean lead concentration for the study area. The formulas for computing the mean and the variance of the estimate mean for both balanced ranked set sampling (as in this example) and unbalanced ranked set sampling designs are provided below.

How do you compute the mean and variance of the estimated mean when balanced ranked set sampling is used?

The true mean of the population is estimated by computing the arithmetic mean of the n measurements obtained on the n samples obtained using a balanced ranked set sampling. The formula is:

$$\bar{x}_{RSS,balanced} = (1/rm)\sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij} \qquad 7.4A$$

where

r m = n

= total number of ranked set sampling samples obtained using a balanced ranked set sampling design

x_{ij} = the measurement of the sample collected from the field location that had rank I that was collected in the j^{th} cycle of sampling

The variance of the estimated mean $\bar{x}_{RSS,balanced}$ is computed as follows:

$$\text{Var}\,(\bar{x}_{RSS,balanced}) = \sum_{i=1}^{m}\sum_{j=1}^{r}(x_{ij} - \bar{x}_i)^2 / m^2 r(r-1) \qquad 7.5A$$

where x_{ij} was defined above and

$\bar{x}_i$ = the arithmetic mean of the r measurements of the r samples from field locations that had rank I collected during the r cycles of sampling.

$$= (1/r)\sum_{j=1}^{r} x_{ij} \qquad 7.6A$$

The standard deviation of $\bar{x}_{RSS,balanced}$ is the square root of Equation 7.5A.

Sample Size for an Unbalanced Ranked Set Sampling Design

The same two-step process described in this appendix is used to develop either a balanced or an unbalanced design: that is, to first determine if ranked set sampling is expected to be more cost effective than simple random sampling, and if so, then determine the number of samples to be collected. Although more research is needed to develop an optimal method to design an unbalanced ranked set sampling design, the "t-model" method developed by Kaul et al. (1995, pp. 6-20) appears to be a reasonable approach that should be satisfactory in practice. Recall that we should consider using an unbalanced design if the distribution of the laboratory measurements is expected to be skewed to the right.

The "t-model" method is to collect *r* samples for each of the *m*-1 smallest ranks (*m* = set size) and to collect *r* x *t* samples for the largest rank, where *t* is some integer greater than 1. For example, if the set size is *m* = 3 and the number of cycles is *r* = 5, a balanced ranked set sampling design results in collecting a sample at each of five locations expected to have a relatively small value of the variable of interest (e.g., lead) as well as at 5 locations that have a mid-value of lead and at five locations that have a relatively large value of lead. But for an unbalanced design, Kaul et al. (1995) suggest collecting a sample at 5 x *t* locations rather than five locations that are expected to have relatively large values of lead. However, the number of locations samples that are expected to have low or middle values of lead remain unchanged at 5. If the optimal value of *t* is selected, then the relative precision of the unbalanced ranked set sampling design is greater than the relative precision of the balanced ranked set sampling design.

Optimum values of *t* for various values of the coefficient of variation (CV) for set sizes of *m* = 2, 3, 4 and 5 are plotted in Figure 6 of Kaul et al. (1995, p. 14). The curves are essentially identical for these values of *m*. Their results are summarized in Table 22.

Table 22. Optimal Values of t for Determining the Number of Samples Required for an Unbalanced Ranked Set Sampling Design Using the "t-model" of Kaul et al. (1995)

CV	0.25	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
t	1	2	3	5	6	7	8	9	10

The total number of samples collected when the "t-model" method is used is computed as follows: $n = (m - 1 + t)\, r$, where *m* is the prespecified set size, *r* is the number of ranked set sampling cycles used, and *t* is determined from Table 22. The formula for computing *r* is given by Equation 7.3A, the same equation used for a balanced ranked set sampling design. However, the values of the relative precision used in Equation 7.3A will be too small if they are obtained from Table 21 since those relative precision values apply to a balanced ranked set sampling design. In order to approximately correct for this bias, the values of relative precision in Table 21 should be multiplied by the correction factors in Table 23. The corrected relative precision values can then be used in Equation 7.3A to determine the approximate *r*. The correction factors in Table 23 are the approximate percent increase in the relative precision that occurs when an unbalanced "t-model" ranked set sampling design is used instead of a balanced design.

Table 23. Correction Factors* for Obtaining Relative Precision Values

CV	0.1	0.3	0.5	0.8	1.3
Correction Factor	1.01	1.08	1.2	1.5	1.7

*Multiply the relative precision in Table 21 by these correction factors to obtain the approximate relative precision value to use in Equation 7-3A to determine the number of ranked set sampling cycles, r.

Example

This example illustrates the above description of how to determine the number of samples required to estimate the mean when an unbalanced design is used. This will be done using the lead contamination example that was used above to illustrate a balanced ranked set sampling design. Recall that the correlation between laboratory lead measurements and the XRF in-site detector readings of lead was approximately 0.9. Furthermore, the cost ratio (the cost of a laboratory lead measurement divided by the cost of using XRF measurements to rank a field location) was approximately 10. Hence, as the cost ratio 10 is larger than any of the cost ratios in Table 20 for any choice of m (set size) when the correlation is 0.9, ranked set sampling is expected to be more cost effective than simple random sampling for any value of m that might be used. Therefore, we proceed to use the same five-step process used above for the balanced ranked set sampling design to determine the number of samples that should be collected and measured for lead in the laboratory.

Step 1: Determine the number of field samples, n_o, needed to estimate the true mean assuming that simple random sampling is used to identify the locations where samples will be collected.

One approach for determining n_o is to use the method given in Table 6 (Chapter 4). However, that method assumes that n_o is large enough such that the estimated mean, $\overline{x}$, will have a normal distribution. However, if the distribution of the laboratory lead measurements is highly skewed (very asymmetrical), then the distribution of $\overline{x}$ may not be normal, in which case the method in Table 6 of Chapter 4 (illustrated in Table 8) may provide a value of n that is too small. One way to mitigate this effect is to use a conservative (i.e., too large) value of the CV in the sample size formula, which will result in a larger value of n_o. Another approach, since the lead measurements are expected to be skewed (or else a balanced ranked set sampling design would be used rather than an unbalanced ranked set sampling design), is to use the method described in Section 3.6 of Perez and Lefante (1997) to determine the number of samples for estimating the mean of a *lognormal* distribution. Of course, if the distribution is skewed but not lognormal, then the number of samples obtained using their method may be too small or too large. The method in Table 6 of Chapter 4 is recommended for general use unless there is high confidence that the distribution is truly lognormal.

Suppose that past studies on similar residence properties had produced 50 lead measurements of the same type as will be obtained in the present study. Graphical displays of the data (probability plots, histograms and box plots) and statistical hypothesis tests for distribution shape indicated that the data were skewed to the right, but not necessarily lognormal. Also, the CV computed using the measurements was 0.7, but this value was increased to 1.0 to help assure that the computed number of samples needed is not too small. Furthermore, during the

DQO process it was decided that the percent relative error of the estimated mean should be no more than 25 percent. Hence, from Table 8 in Chapter 4 the number of samples required is determined to be approximately $n_o = 64$.

Step 2: Select the set size, m.
Suppose the set size m was selected to be m = 5.

Step 3: Determine the value of the multiplicative factor t.
For CV = 1.0, Table 22 shows that t = 3.

Step 4: Determine the value of r, the number of cycles of ranked set sampling for the m-1 = 5-1 = 4 smallest ranks.

Equation 7.3A is used to determine r:

$$r = (n_o / m) \times (1 / RP) \qquad 7.3A$$

We know from steps 1 and 2 above that $n_o = 64$ and m = 5. Now, using Table 21 with m = 5 and CV = 1.0, the value of the relative precision obtained is 1.84 (using linear interpolation in the table). However, this value of relative precision must be increased to correct for the (expected) skewness of the lead data set. The multiplicative correction factor is approximately 1.58, which is obtained by entering Table 23 with CV = 1.0 and using linear interpolation. Therefore the correct value of relative precision is

$$RP = 1.84 \times 1.58 = 2.91$$

Therefore,

$$r = (n_o / m) \times (1 / RP) = (64 / 5) \times (1 / 2.91) = 4.4 \text{ or } 5.$$

Step 5. Compute the total number of samples.

The total number of samples to be collected for the unbalanced ranked set sampling design is

$$n = (m + t - 1) r = (5 + 3 - 1) 5 = 35 ,$$

and r = 5 soil samples will be collected for each of the m - 1 = 5 - 1 = 4 smallest ranks and r x t = 5 x 3 = 15 samples will be collected for the largest rank (rank 5).

The procedure for actually identifying which n = 35 field locations to sample is as follows:

1. Use simple random sampling to select m + t - 1 = 5 + 3 - 1 = 7 sets of m = 5 field locations.
2. Use the XRF in-situ detector to rank from 1 to 5 the five locations within each set.
3. For the first set of five locations, collect a sample at the location with rank 1.
4. For the second set, collect a sample at the location with rank 2.
5. For the third set, collect a sample at the location with rank 3.
6. For the fourth set, collect a sample at the location with rank 4.
7. For each of the 5th, 6th, and 7th sets, collect a sample at the location with rank 5.

The above procedure yields seven field samples. Now, repeat steps 1 through 7 a total of r = 5 times (cycles) to obtain the n = 35 samples required. Note that r x t = 5 x 3 = 15 soil samples are collected at locations with rank 5 and r = 5 soil samples are collected for each of ranks 1, 2, 3, and 4, as required by the "t-model" design.

How do you compute the mean and variance of the estimated mean when unbalanced ranked set sampling is used?

The true mean of the population is estimated by computing the weighted mean of the n measurements obtained on the n samples obtained using the unbalanced. The formula is:

$$\bar{x}_{RSS,unbalanced} = (1/m)\sum_{i=1}^{m}\left(\sum_{j=1}^{r_i} x_{ij} / r_i\right) \qquad 7.7A$$

The variance of the estimated mean $\bar{x}_{RSS,unbalanced}$ is computed as follows:

$$\text{Var}\left(\bar{x}_{RSS,unbalanced}\right) = (1/m^2)\sum_{i=1}^{m}\left[\sum_{j=1}^{r_i}(x_{ij} - \bar{x}_i)^2 / r_i(1 - r_i)\right] \qquad 7.8A$$

We see that equations 7.7A and 7.8A simplify to the equations 7.4A and 7.5A for the balanced ranked set sampling case if $r_i = r_2 \ldots = r_m = r$, that is if a balanced ranked set sampling design is used. The standard deviation of $\bar{x}_{RSS,unbalanced}$ is the square root of Equation 7.8A.

CHAPTER 8

ADAPTIVE CLUSTER SAMPLING

8.1 OVERVIEW

Adaptive sampling designs are designs in which additional units or sites for observation are selected depending on the interpretation of observations made during the survey. Additional sampling is driven by the observed results from an initial sample. Several designs may be considered adaptive sampling designs. However, this chapter will only discuss adaptive cluster sampling.

Adaptive cluster sampling involves the selection of an initial probability-based sample. Typically, additional samples are selected for observation when a characteristic of interest is present in an initial unit or when the initial unit has an observed value meeting some pre-specified condition (e.g., when a critical threshold is exceeded). Choosing an adaptive cluster sampling design has two key elements: (1) choosing an initial sample of units and (2) choosing a rule or condition for determining adjacent units to be added.

Adaptive cluster sampling is considered in situations where the characteristic of interest is sparsely distributed but highly aggregated. Examples of such populations can be found in fisheries (shrimps clustering in large but scattered schools), mineral investigations (unevenly distributed ore concentrations), animal and plant populations (rare and endangered species), pollution concentrations and hot spot investigations, and epidemiology of rare diseases. Adaptive cluster sampling is most useful when quick turnaround of analytical results is possible (e.g., with the use of field measurement technologies). Possible environmental applications of adaptive cluster sampling include soil remediation (investigating the extent of soil contamination), hazardous waste site characterizations, surveying Brownfields, and determining the extent of occurrence of effects of an airborne source of pollutant on nearby flora and fauna.

8.2 UNDER WHAT CONDITIONS IS ADAPTIVE CLUSTER SAMPLING APPROPRIATE?

Consider the following scenario for a contamination study. In most places sampled, contamination is light or negligible, but a few scattered pockets of high contamination are encountered. There are two questions of interest. First, what is the average level of contamination for the whole area? Second, where are the hot spots located? Using the traditional statistical approach, a random or systematic sample of sites or units would be selected and the contaminant measured at each selected site. The average of these measurements provides an unbiased estimate of the population average. The individual observations can be used to create a contour map to locate peaks of contamination. However, the traditional statistical approach has problems. If contamination is negligible over most of the area, the majority of the measurements will be zero or non-detects. Further, random sampling may miss most of the pockets of higher

concentration. Thus, the sample average, as an unbiased estimator of the population mean, will be associated with higher variability. The contour map may not be as accurate in the areas of higher contamination levels.

Adaptive cluster sampling could provide a better approach in situations similar to the one described above. For populations where the characteristic of interest is sparsely distributed but highly aggregated, adaptive cluster sampling can produce substantial gains in precision (lower variability) over traditional designs using the same sample sizes.

8.3 WHAT ARE THE BENEFITS OF USING ADAPTIVE CLUSTER SAMPLING?

Adaptive cluster sampling has several benefits. First, unlike traditional designs which focus only on one objective, it addresses both the objective of estimating the mean concentration and the objective of determining the extent of contamination. Adaptive cluster sampling concentrates resources in areas of greater interest. In a hot spot investigation, for instance, interest is on areas with high levels of contamination. Adaptive cluster sampling directs selection of additional sampling units to these high contamination areas, provided that the initial sample "hits" the areas of interest.

In addition, field technologies used in adaptive cluster sampling can provide quick turnaround time on test results and allow fewer sampling events. Finally, additional characteristics can be observed, adding to the overall usefulness of the study. For instance, in presence/absence studies of rare animal populations, measurements on size, weight, and so forth can be made on the animals that are found.

8.4 WHAT ARE THE LIMITATIONS OF USING ADAPTIVE CLUSTER SAMPLING?

The iterative nature of adaptive cluster sampling introduces several limitations. With adaptive cluster sampling, the process of sampling, testing, resampling and testing may require considerable time. If quick and cheap field measurements are not readily available, the total sampling costs could quickly add up. Because the sampling process stops only when no more units are found to have the characteristic of interest, the final overall sample size is a random quantity. This makes the total cost also a random quantity. Although it is possible to budget for the sampling process using expected total cost; the expected total cost also depends strongly on the validity of the assumption that the characteristic of interest is not widely spread. Consider a contamination investigation where only a few small areas of high contamination are assumed. Suppose this assumption is not valid; that is, the contamination is more widespread, almost throughout the entire study area. The initial sample has a high probability of "hitting" a contaminated area. Because the contaminated areas are widespread, the follow-up sample size will be larger, so the total sample size will be closer to the number of sampling units in the population, resulting in a much higher total cost.

Adaptive cluster sampling is also limited in terms of statistical theory and analytical methods available beyond estimating the mean and variance. The usual sample average and sample variance are unbiased only if the initial probability-based sample is used. The Appendix to this chapter discusses some unbiased estimators of the mean and variance using the entire adaptive cluster sample. Current statistical studies are being made to obtain readily usable inferential tools (confidence intervals, hypothesis testing, etc.) that have been modified for adaptive cluster sampling. At this point only the initial probability-based sample can be readily used for statistical inference. The user needs to be extremely careful when making inferences about the entire study area based only on the initial sample.

Table 24 summarizes some comparisons of features between simple random sampling or grid sampling and adaptive cluster sampling with an initial simple random sample or grid sample.

Table 24. Comparison of Designs

Feature	**Conventional Sampling**		**Adaptive Cluster Sampling**	
	Simple Random Sample	**Grid Sample**	**With Initial Simple Random Sample**	**With Initial Grid Sample**
Unbiased estimators for mean and variance?	Yes	Yes	Yes	Yes
Confidence limits/hypothesis tests?	Yes	Yes	Yes[1]	Yes[1]
Quantifiable decision error rates?	Yes	Yes	Yes[1]	Yes[1]
Hot spot detection probabilities?	No	Yes	No	Yes[1]
Extent of detected hot spots?	No	No	Yes	Yes
Sample size computations feasible?	Yes	Yes	No	No
Sampling cost prediction feasible?	Yes	Yes	No	No

[1]Only based on initial sample size

8.5 HOW DO YOU IMPLEMENT ADAPTIVE CLUSTER SAMPLING?

Adaptive cluster sampling design is implemented using the following basic elements : (1) selecting the initial probability-based sample, (2) specifying a rule or criterion for performing additional sampling, and (3) defining the neighborhood of a sampling unit.

A grid is placed over a geographical area of interest (target population) where each grid square is a potential (primary) sampling unit. This is illustrated in Figure 23(a). Shaded areas on the figure indicate the area of interest; for instance, areas of elevated contaminant levels. This example has three regions of contamination. The 10 darkened rectangles in the figure represent a randomly selected set of 10 sampling units constituting the initial sample. Whenever a sampled unit is found to exhibit the characteristic of interest — that is, the unit intersects any part of the shaded areas — neighboring sampling units are also sampled using a consistent pattern. An example follow-up sampling pattern is shown in Figure 24, where the x's indicate the neighboring sampling units to be sampled. The follow-up sampling pattern is called the *neighborhood* of a sampling unit. The five grid units in the figure make up the neighborhood of the initially sampled unit. In Figure 23(a), three initial sampling units intersect the shaded areas. The units adjacent to these three initial units are sampled next, as shown in Figure 23(b). Some of these sampled adjacent units also intersect the shaded areas, so the units adjacent to these are sampled next, as shown in Figure 23(c). Figures 23(d) to (f) show subsequent sampling until no more sampled units intersect the shaded areas. Figure 25 shows the initial random sample and the final sample. Note that the final sample covers two of the three regions of contamination. If at least one of the initial units had intersected the third area, it would also have been covered by a cluster of observed units.

The final sample consists of *clusters* of selected (observed) units around the initial observed units. Each cluster is bounded by a set of observed units that do not exhibit the characteristic of interest. These are called *edge units*. A cluster without its edge units is called a *network*. Any observed unit, including an edge unit, that does not exhibit the characteristic of interest is a network of size one. Hence, the final sample can be partitioned into non-overlapping networks. These definitions are important in understanding the estimators for statistical parameters like the mean and variance discussed in the appendix to this chapter.

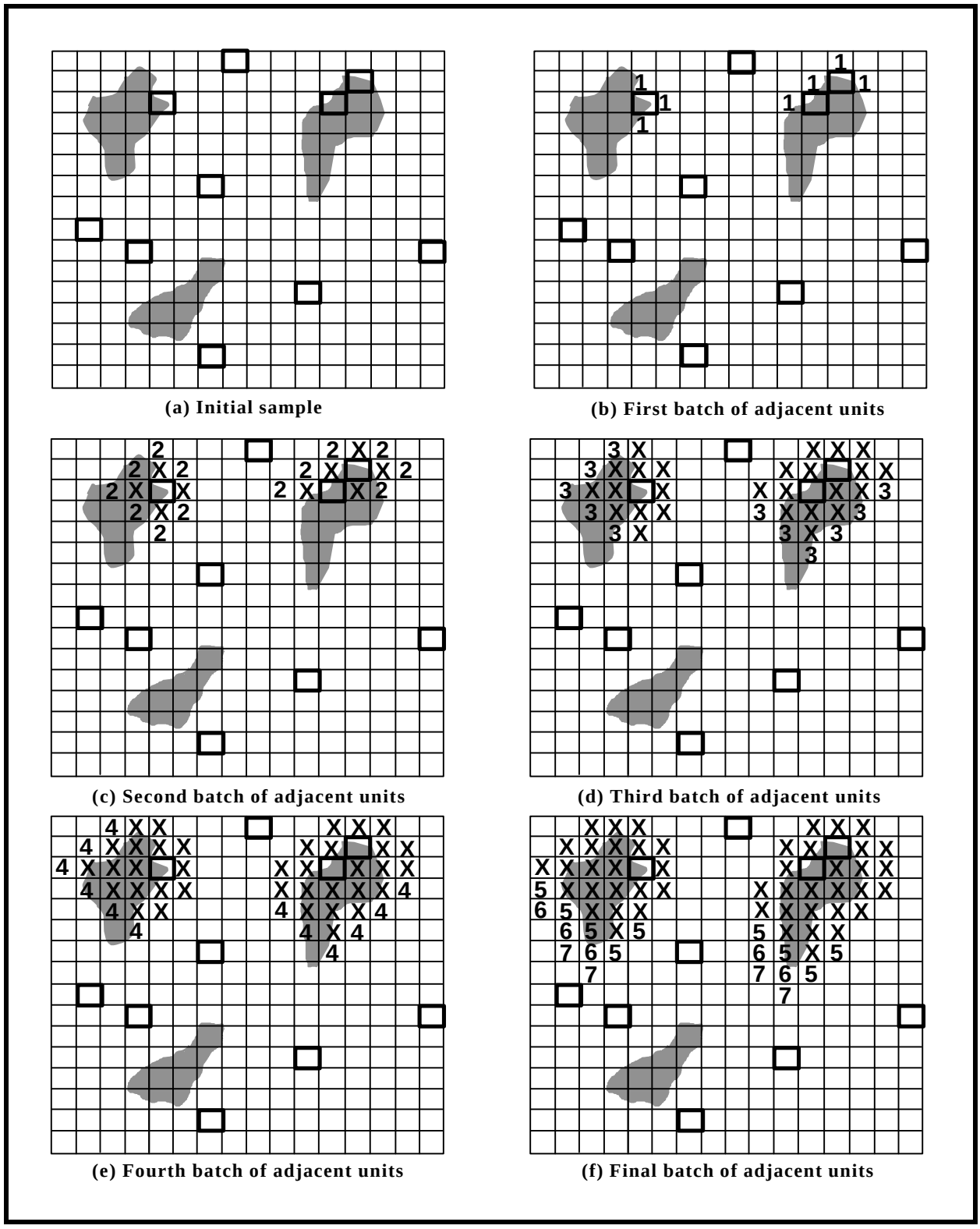


Figure 23. Population Grid with Shaded Areas of Interest, Initial Simple Random Sample, and Follow-up Sample

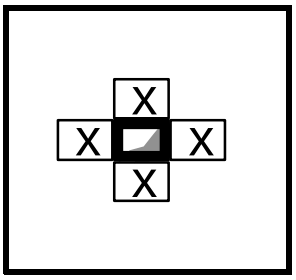


Figure 24. Follow-up Sampling Pattern

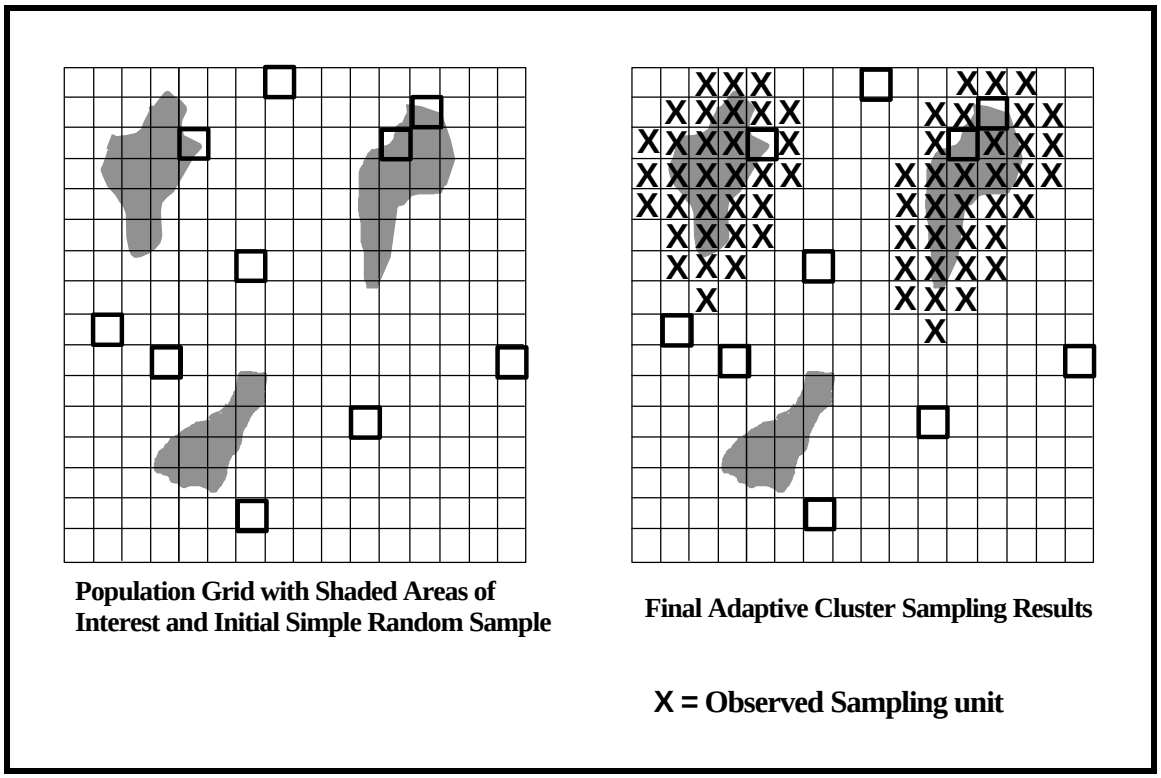


Figure 25. Population Grid with Shaded Areas of Interest, Initial Simple Random Sample, and Final Sample

8.6 UNDER WHAT CONDITIONS ARE OTHER DESIGNS COMMONLY USED IN CONJUNCTION WITH ADAPTIVE CLUSTER SAMPLING?

The initial sample may be obtained using a number of traditional sampling designs. The choice of an initial sampling design is based on the available information about the distribution of the characteristic of interest: possible locations of aggregations or clusters, patterns of contamination, direction of contamination. If little is known of the extent or distribution of the

characteristic of interest over the study region, an initial simple random sample may be useful. If prior information is available, then a stratified sampling or grid sampling approach can be utilized.

An alternative scheme uses primary and secondary sampling units. Suppose the study area in Figure 23(a) is divided into vertical rectangular strips, each one square wide. The strips are the primary sampling units, each of which consists of secondary sampling units (squares). An initial random sample of strips is obtained. If any one of the secondary units within a sampled strip is found to have the characteristic of interest, then the neighborhood of that secondary unit is sampled (the neighborhood is defined as in Figure 24). This scheme is particularly useful when sampling large areas. Seber and Thompson (1994) provide illustrations of these alternative schemes.

8.7 EXAMPLE

Overflow from an impoundment containing nuclear liquid waste historically flowed into an adjacent field (Figure 26). In the field, the flow separated into multiple distinct channels before discharging into a river. The outflow has been shut off for 10 years, the field has been paved into a parking lot, and a new building has been proposed for that parking lot. There is no available information indicating the former locations of the flow channels (e.g., no aerial photographs or surveys). The contamination distribution needs to be characterized to evaluate the potential for contamination to migrate into the adjacent river. Construction of a new building could expose workers and future building inhabitants to contamination. Considering that the contamination is likely to be clustered within the former flow channels and that little prior information is known about the specific locations of the channels (the area cannot easily be stratified), adaptive cluster sampling is an ideal sampling design for this situation. A grid would be established across the parking lot. An initial random sample would be collected, and wherever the concentration or radioactivity exceeds threshold values, neighboring locations would be sampled. Neighboring locations would be sampled in an iterative process until the entire distribution of contamination in the field is characterized. The initial sample could either be a simple random sample, a grid sample, or a strip sample with the strips oriented horizontally from the impoundment towards the river.

To obtain an initial strip sample, first divide the study region into horizontal strips of equal length and width. The region of interest here is the parking lot. Next, divide each strip into smaller, equal-sized quadrants. Select a simple random sample of strips. Each smaller unit in the selected strips is observed. If a smaller unit satisfies condition C, neighboring smaller units are further observed. Although this initial sample design stands the best chance of capturing most of the contaminated areas, it also has the potential for a very large final sample size if the contamination is really not clustered around the former flow channels as assumed. If the contamination is more wide-spread, an initial simple random sample or grid sample may lead to a more cost-efficient final sample. Numerical examples in Appendix 8-A show how to calculate estimates of the mean and variance when using an adaptive cluster sampling design with an initial simple random sample.

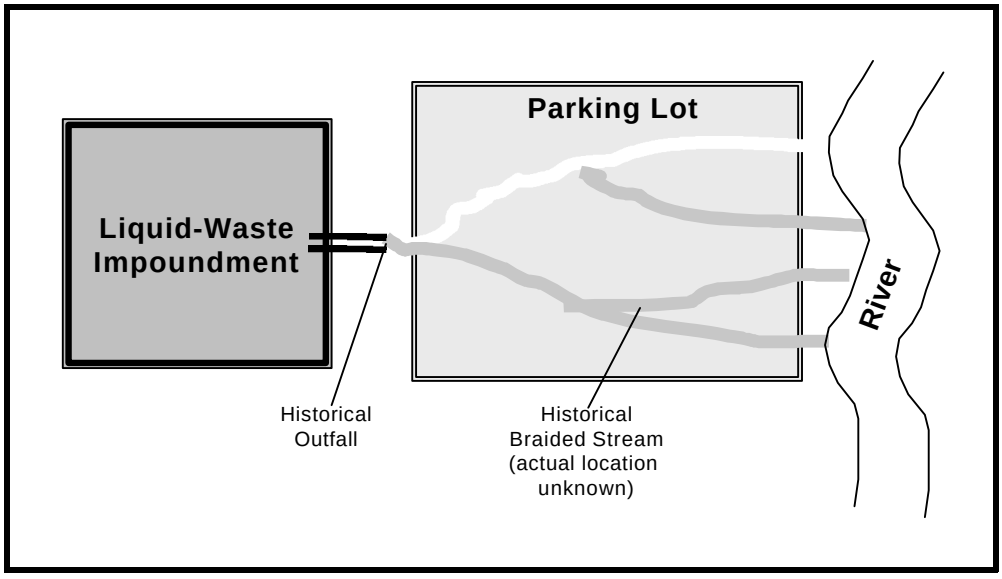


Figure 26. Illustration of an Ideal Situation for Adaptive Cluster Sampling

APPENDIX 8-A

ESTIMATORS OF MEAN AND VARIANCE FOR ADAPTIVE CLUSTER SAMPLING WITH AN INITIAL SIMPLE RANDOM SAMPLE

Select a simple random sample of size n_1 from a population of N units (e.g., grid units or strip units). Define and determine the neighborhood of each unit. An example of a neighborhood would be an initial unit and the immediately adjacent units forming a cross (see Figure 24). For each unit, say I, in the initial sample, determine whether or not the character of interest satisfies a specified condition C (e.g., $y>c$). If so, sample all the units in the neighborhood of unit I. If any of the units in the neighborhood of unit I satisfy condition C, then sample the neighborhood(s) of these units as well. Continue sampling neighborhood units until no more units satisfy the condition. The clusters of units in the sample are bounded by edge units, or units that do not satisfy the condition but are either included in the initial sample or are in the sampled neighborhoods in the follow-up sample. The units in each cluster that are not edge units form a network. Any observed unit, including edge units, that does not satisfy C is considered to be a network of size one. This sampling design partitions the N population units into distinct and disjoint networks.

Note that the method used for obtaining physical samples for analysis within each sample unit depends on the type of application. In particular, for environmental applications, study objectives and decision rules would determine if a single soil grab sample within each unit is adequate or whether or not a composite sample, obtained by combining soil from throughout the unit, should be used.

The usual sample average and sample variance (from a simple random sample) are going to be biased when calculated using the entire final sample. If only the initial sample is used for estimating the mean and variance, unbiased estimators based on the initial sample design can be obtained.

Thompson (1990) has investigated unbiased estimators of the mean and variance based on the final sample. For an adaptive cluster sample with an initial simple random sample, the modified Horvitz-Thompson form of the estimators are

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k} \tag{8A-1}$$

and

$$\hat{\text{var}}(\hat{\mu}) = \frac{1}{N^2} \left[\sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa} \frac{y_j^* \, y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right] \tag{8A-2}$$

where y_k*= sum of the values of the character of interest, y, for the k^{th} network in the sample
N = number of units in the population
k = number of distinct networks in the sample
a_k = probability that the initial sample intersects the k^{th} network
a_{jk} = probability that the initial sample intersects both the j^{th} and the k^{th} networks

Units in the initial sample that do not satisfy the condition C are included in the calculation as networks of size one, but edge units are excluded.

If there are x_k units in the kth network, then the intersection probabilities α_k and α_{jk} are calculated using combinatorial formulas as follows:

$$\alpha_k = 1 - \left[\binom{N-x_k}{n_1} \Big/ \binom{N}{n_1}\right] \tag{8A-3}$$

$$\alpha_{jk} = 1 - \left[\binom{N-x_j}{n_1} + \binom{N-x_k}{n_1} - \binom{N-x_j-x_k}{n_1}\right] \Big/ \binom{N}{n_1} \tag{8A-4}$$

where $a_{jj} = a_j$

Example

Consider the adaptive cluster sample shown in Figure 25. There are N=256 grid units in the population and n_1=10 units in the initial sample. One initial sample unit on the upper left area of the study region intersected a network of x_1=18 units. Let this be network A_1. Two other initial sample units on the upper right area of the study region intersected a network (A_2) of x_2=19 units. The remaining seven initial sample units form networks of size one (A_3, A_4,...,A_9). Hence, there are k=9 distinct networks, with x_1=18, x_2=19, $x_3 = x_4 =...= x_9 =1$ units, respectively. The intersection probability for network A_1 is

$$\alpha_1 = 1 - \left[\binom{256-18}{10} \Big/ \binom{256}{10}\right] = 1 - \left[\left(\frac{238!}{10!228!}\right) \Big/ \left(\frac{256!}{10!246!}\right)\right] = 0.5241791$$

while the intersection probability for network A_2 is

$$\alpha_2 = 1 - \left[\binom{256-19}{10} \Big/ \binom{256}{10}\right] = 1 - \left[\left(\frac{237!}{10!227!}\right) \Big/ \left(\frac{256!}{10!246!}\right)\right] = 0.5441714$$

For the remaining networks (A_3, A_4,...,A_9) the intersection probability is

$$\alpha_k = 1 - \left[\binom{256-1}{10} \Big/ \binom{256}{10}\right] = 1 - \left[\left(\frac{255!}{10!245!}\right) \Big/ \left(\frac{256!}{10!246!}\right)\right] = \frac{10}{256} = 0.0390625$$

Hence, the estimate of the mean using the Horvitz-Thompson estimator is

$$\hat{\mu} = \frac{1}{256}\left[\frac{y_1^*}{0.5241791} + \frac{y_2^*}{0.5441714} + \frac{(y_3^* + y_4^* + \ldots + y_9^*)}{0.0390625}\right]$$

where y_1^* is the sum of the 18 observations from network A_1, y_2^* is the sum of the 19 observations from network A_2, and y_3^*, y_4^*, ..., y_9^* are the single observations from the networks of size one. To compute an estimate of the variance, we need the joint intersection probabilities:

$$\begin{aligned}
\alpha_{12} &= \alpha_{21} \\
&= 1 - \left[\binom{256-18}{10} + \binom{256-19}{10} - \binom{256-18-19}{10}\right] \Big/ \binom{256}{10} \\
&= 1 - \left[\left(\frac{238!}{10!228!}\right) + \left(\frac{237!}{10!227!}\right) - \left(\frac{219!}{10!209!}\right)\right] \Big/ \left(\frac{256!}{10!246!}\right) \\
&= 0.2719547
\end{aligned}$$

$$\begin{aligned}
\alpha_{1j} &= \alpha_{j1} \quad \text{for } j = 3, 4, \ldots, 9 \\
&= 1 - \left[\binom{256-18}{10} + \binom{256-1}{10} - \binom{256-19}{10}\right] \Big/ \binom{256}{10} \\
&= 1 - \left[\left(\frac{238!}{10!228!}\right) + \left(\frac{255!}{10!245!}\right) - \left(\frac{237!}{10!227!}\right)\right] \Big/ \left(\frac{256!}{10!246!}\right) \\
&= 0.0190701
\end{aligned}$$

$$\begin{aligned}
\alpha_{2j} &= \alpha_{j2} \quad \text{for } j = 3, 4, \ldots, 9 \\
&= 1 - \left[\binom{256-19}{10} + \binom{256-1}{10} - \binom{256-20}{10}\right] \Big/ \binom{256}{10} \\
&= 1 - \left[\left(\frac{237!}{10!227!}\right) + \left(\frac{255!}{10!245!}\right) - \left(\frac{236!}{10!226!}\right)\right] \Big/ \left(\frac{256!}{10!246!}\right) \\
&= 0.0198292
\end{aligned}$$

$$
\begin{aligned}
\alpha_{jk} &= \alpha_{kj} \quad \text{for } j = 3, 4, \ldots, 9 \text{ and } k = 3, 4, \ldots, 9,\ j \neq k \\
&= 1 - \left[\binom{256-1}{10} + \binom{256-1}{10} - \binom{256-2}{10} \right] \Big/ \binom{256}{10} \\
&= 1 - \left[\left(\frac{255!}{10!245!} \right) + \left(\frac{255!}{10!245!} \right) - \left(\frac{254!}{10!244!} \right) \right] \Big/ \left(\frac{256!}{10!246!} \right) \\
&= 0.0013786
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\hat{\text{var}}(\hat{\mu}) \\
&= \frac{1}{(256)^2} \left[\sum_{j=1}^{9} \sum_{k=1}^{9} \frac{y_j^* \, y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right] \\
&= \frac{1}{(256)^2} \left[\sum_{j=1}^{9} \frac{(y_j^*)^2}{\alpha_j} \left(\frac{1}{\alpha_j} - 1 \right) + 2 \sum_{j=1}^{8} \sum_{k=j+1}^{9} \frac{y_j^* \, y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right] \\
&= \frac{1}{(256)^2} \left[\frac{(y_1^*)^2}{\alpha_1} \left(\frac{1}{\alpha_1} - 1 \right) + \ldots + \frac{(y_9^*)^2}{\alpha_9} \left(\frac{1}{\alpha_9} - 1 \right) \right] \\
&+ \frac{1}{(256)^2} \left[2 \left(\frac{y_1^* \, y_2^*}{\alpha_{12}} \right) \left(\frac{\alpha_{12}}{\alpha_1 \alpha_2} - 1 \right) + \ldots + 2 \left(\frac{y_8^* \, y_9^*}{\alpha_{89}} \right) \left(\frac{\alpha_{89}}{\alpha_8 \alpha_9} - 1 \right) \right] \\
&= \frac{1}{(256)^2} \left[\sum_{j=1}^{9} \frac{(y_j^*)^2}{\alpha_j} \left(\frac{1}{\alpha_j} - 1 \right) + 2 \sum_{j=1}^{8} \sum_{k=j+1}^{9} \frac{y_j^* \, y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right] \\
&= \frac{1}{(256)^2} \left[\frac{(y_1^*)^2}{0.524179} \left(\frac{1}{0.524179} - 1 \right) + \ldots + \frac{(y_9^*)^2}{0.0390625} \left(\frac{1}{0.0390625} - 1 \right) \right] \\
&+ \frac{1}{(256)^2} \left[2 \left(\frac{y_1^* \, y_2^*}{0.2719547} \right) \left(\frac{0.2719547}{(0.524179)(0.5441714)} - 1 \right) + \ldots + 2 \left(\frac{y_8^* \, y_9^*}{0.0013786} \right) \left(\frac{0.0013786}{(0.0390625)(0.0390625)} - 1 \right) \right]
\end{aligned}
$$

The second type of estimator is a modified Hansen-Hurwitz estimator and is based on the numbers of initial intersections. For an initial simple random sample, the estimators have the form

$$
\tilde{\mu} = \frac{1}{n_1} \sum_{i=1}^{N} \frac{y_i \, f_i}{m_i} = \frac{1}{n_1} \sum_{i=1}^{n_1} w_i = \bar{w} \qquad \text{(8A-5)}
$$

and

$$\hat{\text{var}}(\tilde{\mu}) = \frac{N - n_1}{N n_1 (n_1 - 1)} \sum_{i=1}^{n_1} (w_i - \tilde{\mu})^2 \qquad \textbf{(8A-6)}$$

where y_i = value of the character of interest, *y*, for the i^{th} unit
n_1 = number of units in the initial sample
N = number of units in the sample
f_i = number of units in the initial sample which intersect the network A_i that includes unit *I*
m_i = number of observations in the network A_i that includes unit *I*
$w_i = \frac{1}{m_i} \sum_{j \in A_i} y_j$ = mean of the m_i observations in the network A_i that includes unit *I*

Example

Consider again the adaptive cluster sample shown in Figure 25. For this example, N=256 and n_1=10. There is one initial sample unit in network A_1, two in network A_2, and one each in networks A_3, A_4,...,A_9. Hence, $w_1 = \frac{1}{18} y_1^*$, $w_2 = \frac{1}{19} y_2^*$, and $w_j = y_j^*$ for j=3,4,...,9 . As in the previous example, y_j^* represents the sum of the observations in network A_j. The modified Hansen-Hurwitz estimators of the mean and variance are given by

$$\tilde{\mu} = \frac{1}{10}\left[\frac{1}{18} y_1^* + \frac{1}{19} y_2^* + (y_3^* + \ldots y_9^*)\right]$$

$$\hat{\text{var}}(\tilde{\mu}) = \frac{(256-10)}{256(10)(10-1)}\left[\left(\frac{1}{18} y_1^* - \tilde{\mu}\right)^2 + \left(\frac{1}{19} y_2^* - \tilde{\mu}\right)^2 + (y_3^* - \tilde{\mu})^2 + \ldots (y_9^* - \tilde{\mu})^2\right]$$

Both estimators of the mean, $\hat{\mu}$ and $\tilde{\mu}$ are unbiased for the population mean μ. The estimators of the variances, $\hat{\text{var}}(\hat{\mu})$ and $\hat{\text{var}}(\tilde{\mu})$ are also unbiased for $\text{var}(\hat{\mu})$ and $\text{var}(\tilde{\mu})$, respectively. However, $\text{var}(\tilde{\mu})$ tends to be slightly higher than $\text{var}(\hat{\mu})$ (Christman, 2000).

More examples of calculations for these estimators are given in Thompson and Seber (1996).

The relative efficiency of adaptive cluster sampling versus conventional sample designs can be measured by the ratio of the variances of the mean estimators from the designs being compared. Thompson and Seber (1996) discuss several factors that can increase the efficiency of adaptive cluster sampling designs (using the Hansen-Hurwitz estimator $\tilde{\mu}$):

1. When within-network variability is a high proportion of total population variance, indicating clustered or aggregated populations (according to the character of interest).

2. When there is a high degree of geographic rarity of the population, that is, when the number of units is large relative to the number of units satisfying the condition *C*, and the study region is large relative to the area where the contamination levels are high.

3. When the expected final sample size is not much larger than the initial sample size (i.e. when the units satisfying the condition are clustered together in few clusters, and the units not satisfying the condition but included in the sample are also few in number).

4. When units can be observed in clusters, which is less costly than observing the same number of units scattered at random throughout the region.

5. When units observed do not satisfy the condition, which is less costly than observing units that satisfy the condition.

6. When an easy-to-observe auxiliary variable is used to determine additional sampling; this can cut costs by eliminating the need to measure edge units.

Christman (1997) showed that the efficiency of adaptive cluster sampling relative to simple random sampling without replacement also depends on the choice of the condition C *(y>c)* and on the choice of neighborhood. As *c* increases, the within-network variance decreases, and the estimator becomes less efficient (see item 1 above). Also, using a neighborhood structure that does not consider the likely shape of the clusters of rare units may decrease efficiency. For instance, if the rare units tend to be physically adjacent, a neighborhood structure that includes physically adjacent units will tend to be more efficient than a neighborhood structure that does not.

COST MODEL

It is difficult to derived advanced cost estimates for adaptive cluster sampling since sample sizes are random quantities. In some applications, adaptive cluster sampling provides estimates of the population mean with smaller variance (which translates to lower cost for a required degree of precision) than simple random sampling.

Thompson and Seber (1996) use the following cost model for adaptive cluster sampling with n_1 units in the initial sample and ν units in the follow-up sample. The cost components are

C_T = total cost
c_0 = fixed cost, independent of sample sizes (initial or follow-up)
c_1 = marginal cost per unit in the initial sample
c_2 = marginal cost per unit in the follow-up sample

The total cost for a fixed set of initial and follow-up sample units is given by $C_T = c_0 + c_1 n_1 + c_2(\nu - n_1)$. However, since ν is random then the total cost C_T is also random. The expected total cost is given by

$$E(C_T) = c_0 + c_1 n_1 + c_2[E(\nu) - n_1] = c_0 + (c_1 - c_2) n_1 + c_2 E(\nu) \quad \textbf{(8A-6)}$$

In some situations, it costs more to observe a unit that contains information about the character of interest than one that does not. For example, suppose the character of interest is whether a contaminant threshold value is exceeded. If an initial unit is found to have some level of contamination using a quick field measurement, an additional measurement is made to determine the actual level. It would take longer and cost more to observe a unit with a nonzero level than a unit with no contamination. Thompson and Seber (1996) make the point that if comparisons of sampling strategies are to be made on the basis of cost, then "the relative advantage of the adaptive to the nonadaptive strategy would tend to be greater than in comparisons based solely on sample size" (Thompson and Seber, 1996).

OPTIMUM SAMPLE SIZE FOR ESTIMATION OR HYPOTHESIS TESTING

It is difficult, if not impossible, to evaluate the performance of a statistical test of hypothesis when using adaptive cluster sampling. For this reason, it is also difficult or impossible to determine optimal numbers of samples (Step 7 of the Data Quality Objectives Process). In estimating the mean, the final sample size is a random quantity and so cannot be determined in advance. However, in the guidelines proposed by Thompson and Seber (1996) for increasing the efficiency of adaptive cluster sampling (see previous section for a summary of factors), a general idea is provided of how the sample size should be distributed between the edge units and the units satisfying the condition, and how much larger the final sample size should be relative to the initial sample size. Furthermore, if the extent and abundance of the population are underestimated, one could end up with more units than time or cost would allow. If additional sampling is curtailed because of this underestimation, biases can occur in the estimation of the mean. Thompson and Seber (1996) address this issue and offer suggestions on limiting total sampling effort.

CHAPTER 9

COMPOSITE SAMPLING

9.1 OVERVIEW

Composite sampling involves physically combining and homogenizing environmental samples or subsamples to form a new sample (i.e., a composite sample). The chemical or biological analyses of interest are then performed on (aliquots of) the composite sample. Because the compositing physically averages the individual samples, averaging the analytical results of a few composites can produce an estimated mean that is as precise as one based on many more individual sample results. Since fewer analyses are needed, composite sampling can substantially reduce study costs when analysis costs are high relative to the costs associated with the collection, handling, and compositing of the samples. Depending on the particular situation, the particular units that comprise a composite sample may or may not need to have resulted from a prescribed sampling design (e.g., random or grid sampling). Unlike the other sampling approaches described in this guidance, composite sampling does not refer to a statistically based strategy for selecting samples. Although composite sampling may assume that the samples have been collected according to such a design, it further requires that the compositing be carried out according to some protocol (i.e., a composite sampling protocol). The composite sampling protocol identifies how the compositing is to be carried out (e.g., which samples are used to form each composite) and how many composite samples are to be formed. For example, Figure 27 shows a situation in which n=9 individual samples are selected and are used to form m=3 composites of k=3 samples each; hence only m=3 analyses, one per composite, are required.

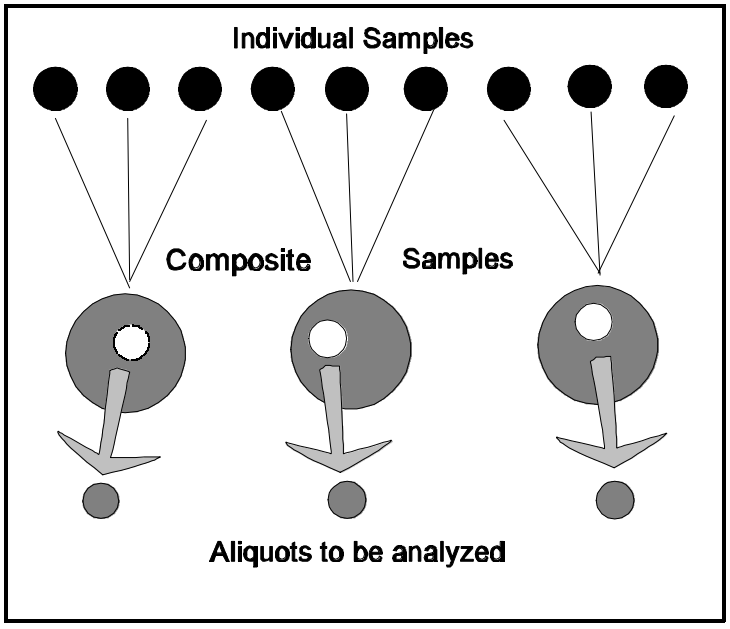


Figure 27. Equal Volume, Equal Allocation Compositing

In the designs considered elsewhere in this document, there is a one-to-one correspondence between the number of samples taken and the number of laboratory analyses performed. In cases where composite sampling can be used, however, there are n samples taken, but only $m<n$ analyses are generally performed, one for each composite. There are several fundamental ideas associated with composite sampling:

- A primary goal is to reduce cost by having fewer analyses.
- Composite sampling requires that the sample acquisition and handling process can be separated from the measurement process.
- Compositing results in a physical averaging of the samples making up the composite so that
  - **S** If the concentrations of a contaminant could be measured accurately in the individual samples as well as in their composite, and if the compositing process is carried out properly, then we would expect the measured level for the composite sample to be equal to the average of the measurements made on the individual samples (assuming no measurement errors).
  - **S** Variability among similarly formed composite samples is less than the variability of the individual samples.
  - **S** Composite sampling is naturally compatible with a study goal of estimating a population mean, while other goals may not be compatible with composite sampling, since some information is lost.

Although composite sampling has historically been used mainly for estimating a mean, it can also be used in some cases to estimate the proportion of a population that has a particular trait.

Another use for composite sampling is in the identification of a rare trait—that is, in classifying samples as having or not having a trait. This requires that only aliquots of the individual samples be composited so that some of the individual samples can be retested, based on the analytical results found for the composite. If any unit comprising a composite has the trait, then the composite will as well. This situation requires a composite sampling and retesting protocol. Since composite sampling and retesting is aimed at classifying each unit, as opposed to making a statistical inference about the population represented by the sample units, composite sampling in this context will typically be used when there is a finite number of units needed to be classified. Suppose, for example, that water samples have been collected from all the drinking water wells within some region, and it is suspected that a small proportion of these *n* wells are contaminated (in some specific way). Rather than testing each individual sample (*n* tests), it may be adequate to group the samples into sets of eight. Within each group, aliquots of equal volume are taken and a test is made on the composite. If the test shows contamination, then aliquots from the individual samples are tested to identify the specific ones that are contaminated; if a small proportion of wells are contaminated, then many of the composites will yield negative results. Thus, such a strategy, or even more sophisticated ones, can substantially reduce the amount of testing. This use of composite sampling and retesting is sometimes called group testing or screening.

In some circumstances, composite sampling can be coupled with retesting to identify the particular unit that has the highest level of a contaminant or the set of units having the highest levels (e.g., those in the upper 1 percent of the distribution). Like the prior situation, this would often be applied when there is a finite number of units of interest. These strategies assume that measurement errors are negligible. The relative magnitudes of the composite sample measurements are used to decide on which composite(s) could potentially contain the unit with the highest level; the individual samples within those composites are then retested to determine which unit has the maximum concentration.

Table 25 identifies the four main situations for which or composite sampling and retesting are meaningfully applied and shows where these are discussed.

Table 25. When to Use Composite Sampling — Four Fundamental Cases

Type	Objective	Section
Objectives that rely on composite sampling	Estimating a population (or stratum) mean for a continuous measure (e.g., analyte concentration)*	9.2
	Estimating proportion of population exhibiting some trait	9.3
Objectives that rely on composite sampling and retesting protocols	Classifying sampling units as having or not having some trait such as a being in a hot spot or from a contaminated well	9.4
	Identifying the sampling unit with highest value of some continuous measure (e.g., concentration), or identifying sampling units in the upper percentiles	9.5

* In general, information on variability and spatial (or temporal) patterns is lost when compositing is used for this objective; however, in some cases, some information on patterns can be acquired.

In the first two cases (Sections 9.2 and 9.3), there is interest in *making an estimate for a prescribed target population*—in the first case estimating the mean of a continuous measure (e.g, the mean concentration of contaminant) and in the second case estimating the proportion of the population with a characteristic. In these two cases, carrying out the CS requires combining a sampling design with a compositing protocol. The sampling design describes the method for selecting units from the target population and indicates the number of units to be selected and which ones are to be selected. The compositing protocol describes the scheme for forming and

processing (mixing and homogenizing) composites. It indicates whether entire samples or aliquots are to be combined, the number of groups of units to be formed (*m*), the number of units per group (*k*), which units form each group, and the amount of material from each unit to be used in forming the composite sample.

The last two cases (Sections 9.4 and 9.5) involve *decision making at the unit level* rather than at the target population level. As a consequence, these approaches involve CS&R protocols that not only define how composites are to be formed but also define when and how subsequent testing is to be done to ultimately identify particular units. The retesting strategies for these cases are conditional on the results obtained for the composite(s). In order to retest individual samples, the identity and integrity of the individual samples must be maintained; this implies that aliquots from the individual samples, rather than the whole samples, must be combined in forming composites. Additional aliquots from the individual samples are then retested either singly or in other composites.

Before considering CS for one of the purposes indicated above, careful consideration needs to be given to its advantages and disadvantages. These are discussed in more detail in the subsequent sections; however, Table 26 provides some general guidance as to when CS may be useful and practical. In addition to its other potential merits, it should be noted that CS may sometimes be needed in order to have an adequate mass for analysis (e.g., for dust samples or tissue samples). Finally, it should be noted that a single investigation may have several objectives —for example, estimating a population mean and its precision, as well as identifying, through retesting, the units with the highest levels. Innovative ways of applying CS need to be considered in these circumstances as a means of achieving major cost savings.

Table 26. Criteria for Judging Benefits of Composite Sampling

Criterion	**CS is likely to be beneficial if...**
Analytical costs	Analytical costs are high relative to sample acquisition/handling costs.
Analytical variability	Analytical variability is small relative to inherent variability.
Objective is to estimate population mean	Information on individual samples is not important. Information on associations is not important.

Table 26. Criteria for Judging Benefits of Composite Sampling

Criterion	CS is likely to be beneficial if...
Objective is to estimate proportion of population with a trait	Composite has trait if individual sample does. Likelihood of misclassification is small. Trait is rare.
Objective is to classify samples as having/ not having a trait	Composite has trait if individual sample does. Likelihood of misclassification is small. Retesting of aliquots from individual samples is possible. Trait is rare.
Objective is to identify the sample(s) with the highest value	Measurement error is negligible. Retesting of aliquots from individual samples is possible.
Range of analytical concentrations	Concentrations of relevance are much larger than detection limits.
Physical barriers	Compositing does not affect sample integrity (expect no chemical reactions/interferences or analyte losses from volatility) or result in safety hazards.
Logistical barriers	Individual samples can be adequately homogenized.

Useful references for understanding compositing and its various uses are Garner et al. (1988), Patil et al. (1996). For estimating a mean, additional material is provided in Chapter 7 of Gilbert (1987).

9.2 COMPOSITE SAMPLING FOR ESTIMATING A MEAN

9.2.1 Overview

This section discusses CS when the objective of the study is to estimate a population mean (e.g., an average site or process concentration of a contaminant). The focus is on the following situation:

1. The individual samples comprising the composite are of equal size (in volume or mass) and shape.

2. The number of samples comprising each composite is the same.

3. A single subsample or aliquot is selected for analysis.

4. A single analysis is performed on the subsample.

5. A large number of composite samples could potentially be formed, but the number to actually be formed is small relative to the number of potential composites.

These are the most common conditions. Condition (1) above is necessary if the composite is to be considered equivalent to a simple averaging of the individual samples. Condition (2) is desirable so that all properties of all the composites will be the same. Conditions (3) and (4) would generally be used when compositing is contemplated, since the variation associated with aliquots and measurements would usually be smaller than that inherent to the population. Gilbert (1987) discusses the modifications to the computations of estimates and statistical analysis that are needed if the units are not of equal size (pp. 79-85) or if conditions (3) or (4) or (5) are relaxed (pp. 72-79).

9.2.2 Under what conditions is composite sampling appropriate?

Composite sampling for estimating a mean will generally be an appropriate strategy when all of the following conditions hold:

1. There are no other goals that conflict with the notion of compositing. In particular, information regarding levels for individual samples, their spatial or temporal locations, and their population variability is not considered important; information on associations (e.g., correlations of concentrations of two constituents) is also not considered important.

2. Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. Otherwise, composite sampling will not be cost-effective.

3. There are no practical difficulties that impede the selection of multiple samples of units, where each sample is selected according to a given statistical design (e.g., a simple random sample or a ranked set sample).

4. There are no practical difficulties in forming appropriate composites (e.g., individual samples can be adequately homogenized).

5. Compositing will not affect the sample integrity.

6. The anticipated levels for most composites will exceed detection limits so that difficulties of mean estimation in the presence of non-detects are avoided.

Two comments concerning condition (1) need to be made. First, since information on variance is lost when samples are composited, repeated application of the composite sampling protocol may be needed to get information on variability that is needed to conduct a hypothesis test or to form a confidence interval. Under these circumstances, one must recognize that there is an implied secondary objective of the study—namely, estimation of the standard error of the estimated mean. Second, some special composite sampling protocols can be employed along with innovative data analysis techniques so that some information on spatial patterns can be resolved. Lancaster and Keller-McNulty (1998), for instance, give an example in which a site is partitioned into rows and columns to form squares and for which composite samples were formed for all squares within each row and for all squares within each column. By using statistical modeling approaches (analysis of variance type concepts) to predict areas likely to have levels of a contaminant (a secondary objective in addition to mean estimation), some information on the spatial patterns was obtained, while estimation of the mean, the primary objective, was also accomplished by appropriately averaging the results from the composite samples.

9.2.3 What are the benefits of using composite sampling?

The main benefit of using composite sampling to estimate a mean is that one can achieve approximately the same precision of an estimated mean at less cost or, alternatively, one can get more coverage (better representation) of the target population at the same (or lower) cost. Box 1 illustrates the substantial benefits that can result.

A second benefit of composite sampling is that data analysis is usually easy. For example, if composite sampling is coupled with simple random sampling or grid sampling (where each composite represents the entire target population), and if samples of equal volumes are used to form the composites, then the simple average of the *m* composite samples provides an estimate of the target population mean.

9.2.4 What are the limitations of using composite sampling?

The main limitations are those implied by the conditions [other than (2)] delineated in Section 9.2.2. As noted, composite sampling yields a reduced amount of information on variability. If the composite sampling protocol allows each composite to represent the entire target population, then the measured concentration of the contaminant provides an estimate of the population mean concentration, assuming the measurement process is unbiased and that the compositing has been carried out properly. However, this process must be repeated several times to estimate the precision of the estimated mean. Composite sampling also loses information on individual samples, spatial (or temporal) patterns, and other associations. For instance, if the compositing is over time, then information on when the contaminant had high concentrations is lost; similarly for space. Temporal and spatial correlations often exist between concentrations of two contaminants, but this information is also lost when composites are used.

Box 1. Example of Benefits of Composite Sampling

Assume zero cost for compositing and no additional variance is introduced due to composting.
Suppose k = 5, m=20 (i.e., 20 composites of 5 units each).
Assume the ratio of analysis to sampling costs is as shown below; then the costs achieved by compositing relative to the cost without compositing (i.e., relative to a simple random sampling of 100 units) is as shown (fixed set-up costs are ignored):

ratio of per-unit analysis cost to per-unit sampling/handling cost:	2	3	4	5	10	20
relative cost of study:	0.47	0.40	0.36	0.33	0.27	0.24

Note that the relative cost approaches 1/k as the ratio gets larger.
Assume the ratio of measurement error standard deviation (σ_M) to inherent variation standard deviation (σ_I) is as shown below; the precision of the estimated mean relative to the mean determined without compositing is as shown:

ratio σ_M/σ_I:	0.01	0.05	0.1	0.2	0.3	0.4
std. error of mean*:	1.000	1.005	1.020	1.074	1.153	1.246

* relative to an simple random sampling of 100 units, without compositing.

Suppose k = 10, m=10 (i.e., 10 composites of 10 units each).
Assume the ratio of analysis to sampling costs is as shown below; then the costs achieved by compositing relative to the cost without compositing (i.e., relative to a simple random sampling of 100 units) is as shown (fixed set-up costs are ignored):

ratio of per-unit analysis cost to per-unit sampling/handling cost:	2	3	4	5	10	20
relative cost of study:	0.40	0.33	0.28	0.25	0.18	0.14

Note that the relative cost approaches 1/k as the ratio gets larger.
Assume the ratio of measurement error standard deviation (σ_M) to inherent variation standard deviation (σ_I) is as shown below; the precision of the estimated mean relative to the mean determined without compositing is as shown:

ratio σ_M/σ_I:	0.01	0.05	0.1	0.2	0.3	0.4

Composite sampling may be difficult to accomplish without introducing large additional errors (in weighing or homogenizing). This is especially true for a solid medium such as soil or dust if they are heterogeneous. Homogeneity in this context refers to characteristics of the environmental medium that affect one's ability to form a "good" composite. In general, Composite sampling works better for more homogeneous media. As Pitard (1993) notes,

however, homogeneity is a relative concept: sand particles seen from a distance may seem homogeneous but the individual grains will reveal great heterogeneity in shape, size, color, density, and so forth. In compositing, one must be assured that the *k* individual samples are mixed to the point that one can expect that 1/*k* of any aliquot came from each individual sample. Achieving this can be especially difficult if the individual samples are dissimilar in properties that affect homogenization. For example, soil composites comprised of some samples that are mostly clay and some that are mostly sand will tend to be poor candidates for composite sampling. On the other hand, liquids tend to be more homogeneous than solids so that liquids are typically good candidates for composite sampling, but even liquids must be thoroughly mixed prior to compositing, if aliquots are to be composited, and prior to taking aliquots from the composite.

Box 2. Directions for Selecting Equal Allocation, Equal Volume Composite Samples for Estimating a Mean

STEP 1:	If compositing costs can be considered negligible relative to other costs and additional variability introduced by the compositing process can be ignored, use Table 27 to determine an initial k value. Determine an appropriate volume for each sample, and determine if the indicated k can be used, based on the physical nature of the samples, the anticipated levels of the contaminant(s) of interest (relative to detection limits), and the capability to combine and homogenize them adequately. If the indicated k is too large based on these practical considerations, use the maximum practical k.
STEP 2:	Choose m so that selection of n = mk samples and analysis of m samples will be within the budget: Cost = fixed set-up costs + cost of selecting/handling n samples + cost of forming m composites + cost of performing m analyses.
STEP 3:	Check that m is large enough to produce a sufficiently precise estimate of the mean: Variance of estimated mean = variance of mean without compositing + measurement error variance*(k-1)/mk (assuming compositing adds no error). If the precision needs to be improved, or if confidence intervals or tests are to be performed for the mean (which may imply the need for a larger m), then consider cost-precision tradeoffs in which m is increased and k is reduced.
STEP 4:	Select k samples according to the prescribed sampling design and compositing protocol. (See Chapter 4 for procedures for selecting a random sample.)
STEP 5:	Form one composite sample as follows: physically mix and homogenize each sample (this may be unnecessary if each sample is of the same volume and each sample is to be included in the composite in its entirety). Obtain equal-volume aliquots of material from each of these k samples and physically mix the selected material and homogenize it thoroughly.
STEP 6:	Repeat steps 4 and 5 m times to form m composite samples.
STEP 7:	Obtain measurements on the environmental parameters of interest for each composite sample or, if necessary, on equal-volume aliquots therefrom.

9.2.5 How do you implement composite sampling?

Box 2 provides directions for implementing composite sampling to estimate a mean in the simplest situation—namely, when equal-volumes and equal-allocations of samples or aliquots are used to form composites and where each composite represents the entire target population (by virtue of random sampling, for instance). Gilbert (1987) and Elder, Thompson, and Myers (1980) describe procedures when equal volumes/allocations are not used.

Steps 1, 2, and 3 involve considerations for arriving at appropriate values for k and m; details are provided in Appendix 9A. The basic mechanism for determining k and m is to develop a cost model that expresses the total cost in terms of its per-unit and per-analysis costs and to develop a variance model that expresses the variance of an estimated mean in terms of variance components:

Total Cost = fixed set-up costs + n(cost of sample collection/handling) + m(cost of analysis)

$$\begin{aligned} \text{Variance (mean)} &= \text{Variance [sum of } m \text{ observed concentrations]}/m^2 \\ &= m \text{ x Variance [observed composite concentration]}/m^2 \\ &= [(\text{inherent concentration variability})/k + \text{error variability}]/m \\ &= [\text{inherent concentration variability} + k\,(\text{error variability})]/n \end{aligned}$$

where $n = mk$

The "inherent concentration variability" component refers to the natural variability in the true concentrations that occurs among units in the target population, whereas the "error variability" component refers to variability resulting from random errors made in the sample collection and measurement processes. To arrive at corresponding formulas for an simple random sampling without compositing, one can substitute m=n and k=1 into the above. The above formulas are simplified since they omit the cost of forming composites, which would typically depend on k, and the additional variance that might result from compositing. Under these simplifying assumptions, an optimal k can be determined as a function of the relative cost components and variance components (see Appendix 9A). A table of optimal k values for this situation is furnished in Table 27. Use of this table requires prior knowledge about the population variability and the anticipated error variability. For instance, suppose a prior study investigating soil contamination resulted in an observed distribution of concentrations of the contaminant of interest that had a coefficient of variation (CV) of 90 percent; based on geological considerations and the fact that the study under consideration plans to use the same measurement methods, a similar CV is anticipated. Analysis of the concentrations from a series of duplicate samples (i.e., quality control data from the prior study) revealed a measurement-error relative standard deviation of 18 percent. The measurement error variance thus accounts for $100(18/90)^2$ percent of the total variance, or 4 percent; hence the inherent variation is anticipated to be 96 percent of the total variance. The ratio of the standard deviation of measurement errors to the standard deviation of inherent levels can therefore be

determined as the square root of (4/96), or approximately 0.20; this identifies the column in Table 27 to be used. Since the per-unit measurement cost is expected to be about three times the cost of sample acquisition, the table shows that using k=9 will be optimal (from a statistical standpoint) for the planned study. As noted above, this assumes both that the compositing of nine samples is practical and that components of cost and variance associated with the compositing process can be ignored.

Table 27. Optimal k Values for Estimating a Population Mean* as a Function of Variability and Cost Components

Cost Ratio C_M/C_S	Ratio of Std. Dev. of Measurement Errors to Std. Dev. of Inherent Levels									
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**1.00**
2	14	7	5	4	3	2	2	2	2	1
3	17	9	6	4	3	3	2	2	2	2
4	20	10	7	5	4	3	3	3	2	2
5	22	11	7	6	4	4	3	3	2	2
8	28	14	9	7	6	5	4	4	3	3
10	32	16	11	8	6	5	5	4	4	3
15	39	19	13	10	8	6	6	5	4	4
20	45	22	15	11	9	7	6	6	5	4
50	71	35	24	18	14	12	10	9	8	7

* The entries assume that compositing costs are negligible, that any increase in variability due to compositing is negligible, and that the samples to be composited are selected at random from the entire target population. C_M and C_S are the per-unit analysis cost and the per-unit sample acquisition/handling cost, respectively.

Box 3 provides an artificial example that demonstrates the benefits of compositing in terms of costs and precision by comparing the results from a particular compositing scheme to those that would have been obtained from a grid sample.

In addition to coupling composite sampling to a simple random sample or a grid sample, several other alternative composite sampling protocols may be worth considering. These include the following:

- If *k* strata of equal size (volume or mass) can be used, then one sample per stratum can be picked and those *k* samples composited. This can avoid having a random sample that is clustered in a small subregion of the target area. If this process is repeated *m* times, then the simple average of the composite sampling results will estimate the target population

Box 3. Example: Compositing for Estimating a Site Mean

The goal was to estimate the average surface soil concentration of contaminant X for a 300' x 400' site. The site was partitioned into 12 square (100' x 100') subareas (cells). Assume that three randomly sited grid samples of 12 points each were placed over the site and that the observed X concentrations for *individual samples* were as follows:

Grid 1:	853	986	1090	2344
	885	1082	767	1592
	528	763	993	806
Grid 2:	983	869	1740	258
	799	756	643	1747
	794	751	985	1106
Grid 3:	1161	886	1256	2276
	838	791	1267	2034
	714	907	913	1118

Data Analysis Results and Costs for Each Grid Sample (and Combined)—Based on Individual Sample Analyses

Grid Number	1	2	3	Combined
Observed Mean Conc.	1057	1147	1180	1128
Std. Error	139	168	142	85
95% Confidence Interval	(752, 1362)	(777, 1517)	(867, 1493)	(960, 1296)
Costs:				
- for collecting/handling	12x$25=$300	12x$25=$300	12x$25=$300	36x$25=$900
- for analysis	12x$500=$6000	12x4500=$6000	12x$500=$6000	36x$500=$18000
- total	$6300	$6300	$6300	$18900

Data Analysis Results and Costs for Each Grid Sample (and Combined)—Based on Composite Sample Analyses

Grid Number	1	2	3	Combined
Observed Mean Conc.*	1070	976	1125	1057
Std. Error				
95% Confidence Interval				(865, 1249)
Costs:				
- for collecting/handling	12x$25=$300	12x$25=$300	12x$25=$300	36x$25=$900
- for compositing	1x$20=$20	1x$20=$20	1x$20=$20	3x$20=$60
- for analysis	1x$500=$500	1x$500=$500	1x$500=$500	3x$500=$1500
- total	$820	$820	$820	$2460

* Assumes the same analytical error variability as for the individual samples; also assumes no additional error occurs as a result of the compositing process.

The above data were generated for a site in which the true site mean was 1113. The true means of the three grids were 1063, 1149, and 1174, respectively, for a true grid average of 1129.

mean, and the variability of the composite sampling results can be used to estimate the precision of the mean. The precision of the mean via this approach may be poorer than that from an simple random sampling, but one is assured that the sample achieves adequate coverage of the target population.

- If compositing within strata is used rather than compositing throughout the site or process, then the precision of the estimated mean may be improved over that achieved by coupling composite sampling with a simple random sample or grid sample from the overall population. This improvement should be realized if the contaminant concentrations are more homogeneous within strata than among strata. A major advantage of this approach is that some spatial or temporal information is attained; also, if retesting is feasible, then information on selected specific units can be efficiently acquired. But a drawback is the difficulty in getting a good estimate of the precision of the mean, especially if the variability within the different strata is different. This drawback occurs because there must be enough composites per stratum to obtain separate variance estimates for each stratum; rather than having a single overall estimate of variance, one in this case forms a weighted combination of the separate estimates to arrive at the estimated precision for the mean.

9.2.6 Under what conditions are other designs commonly used in conjunction with composite sampling?

To produce a meaningful estimate of a population mean, composite sampling must be suitably linked with some type of sampling design. Some examples follow.

Simple random sample. In this case, *m* simple random samples of *k* units each are selected from throughout the site/process, and the units in each sample are used to form composites. In this case, the estimate of the population mean equals the simple average of the *m* composite-sample measurements, if equal-volume samples or aliquots are used. The standard error of the mean (with the same assumption) is calculated based on the variability among the *m* composites. If confidence intervals or hypothesis tests are to be performed (which depend on having a good estimate of the standard error, as well as the mean), then there must be adequate *m* for estimating the variance among the composites. To achieve this, some tradeoff of *m* versus *k* values may be needed.

Grid sample. Composite sampling in this case would involve forming *m* grids having different starting values; each grid would have *k* points, and samples from those points would be composited. The estimation of the mean and its variance would be like that for a simple random sample. See the example in Box 3.

Stratified random or stratified grid sample. If composites are formed from samples within strata, then the target population mean is estimated as a weighted average of the stratum means. The variability of this estimated mean therefore depends on the within-stratum variation. Consequently, we would expect better precision than for a comparably sized random sample, assuming that the concentrations have less variation within strata than among strata. As noted in Section 9.2.5, however, it may be more difficult to get an estimate of the variance (or the standard error) of the mean and the cost may be higher since more composites might be needed to get such a precise estimate. The main advantage of this approach is that it retains some information on spatial or temporal variation.

Two-stage sampling. This is a fairly common situation. In this case, units fall naturally into groups (called "batches") or can be grouped into batches, and h out of H batches are chosen at the first stage of sampling. Then n samples are selected within each batch and m composites are formed from these samples. If $h=H$, then batches are equivalent to strata and this case is like the prior one. But if $h<H$, then this is considered two-stage sampling. For instance, in monitoring a wastewater stream, batches could be days, and samples selected at random times within the selected days could be used to form composites representing days. Analysis of data for the two-stage sampling situation is treated in Gilbert (1987) pages 72-78 for equal sized units and pages 79-85 for unequal sized units. Gilbert also treats the case where multiple subsamples from each composite are extracted and multiple analyses per subsample are performed.

Ranked set sampling. Composite sampling can be used in conjunction with ranked set sampling. Usually it would involve forming composites from units having the same ranks. In this case, the advantages and disadvantages are similar to those for stratified sampling (see above). Compositing across ranks might also be considered. For example, suppose k=3, so that nine samples would be randomly selected and grouped into three groups of three each. The three samples in each group would be ranked by inspection (assumed to be correlated with the parameter of interest). The sample with rank 1 in group 1, the sample in group 2 with rank 2, and the sample in group 3 with rank 3 would be composited and analyzed. In terms of the precision of the estimated mean, such an approach should perform better than a simple random sample of size three (because this is characteristic of ranked set sampling) or a composite therefrom. But it should perform worse than a simple random sample of size nine or a composite of all nine samples. Hence, the compositing across ranks strategy would appear to be a good one only if there were physical or practical reasons why nine samples could not be composited but three could.

9.2.7 Examples

The example in Box 3 illustrates use of composite sampling with a grid sample. Similar approaches would apply if the cells in that example were regarded as strata, since the strata are of equal size. An alternative compositing scheme would involve compositing the samples within each of the 12 strata and averaging the results to produce an estimate of the target population mean. This strategy would be somewhat more expensive and would provide some information on spatial patterns but would not permit one to get a measure of the precision of the overall mean (unless the whole process were replicated). Depending on secondary objectives that may be of interest, other composite sampling strategies might be used that would form composites within rows and within columns.

As noted earlier, compositing sometimes provides samples with mass sufficient to permit use of chemical techniques that allow a broader range of chemicals to be detected. Such protocols may be especially useful for analyzing dust samples (e.g., from various areas within a home) or certain tissue samples (e.g., from several fish).

9.3 COMPOSITE SAMPLING FOR ESTIMATING A POPULATION PROPORTION

9.3.1 Overview

Under certain circumstances, composite sampling offers an efficient way to estimate the proportion of a population that has a particular trait. The goal in this case is simply to estimate the proportion with the trait; there is no interest in which units have the trait. This might be useful in the early stages of investigation when one wants to determine if further testing is warranted. For example, if fewer than 5 percent of the drums in a field are contaminated, then it might be worthwhile to identify them (at a later stage) and selectively remove the contaminated ones, but if more than 15 percent of the drums are contaminated, it may be more cost-effective to simply remove all of them. What is needed as an initial step is an efficient way to estimate the proportion of the drums that are contaminated. In such a situation, *m* random samples of *k* drums could be selected and used to form *m* composites, and analysis of the composites would reveal whether or not they were contaminated. The analysis of each composite in this case simply yields either a positive (i.e., has trait) or a negative (i.e., does not have trait) result. A negative result for a composite implies that all of its component samples were negative (i.e., uncontaminated), while a positive result implies that one or more of the component samples was positive. When composites are formed from random samples, there is a known relationship between the proportion of composite samples expected to be positive and the proportion of individual samples expected to be positive; this relationship allows the proportion of individual samples with the trait to be estimated based on the observed proportion of positive composite samples.

9.3.2 Under what conditions is composite sampling appropriate?

Use of composite sampling for estimating a population proportion, while not a particularly common practice, can be very cost-effective if the conditions for its use are appropriate. All of the following conditions should hold if composite sampling for this purpose is to be used:

- Information on individual samples and/or on their spatial or temporal locations is not considered important.

- Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. Otherwise, composite sampling will not be cost-effective.

- The proportion of the population having the trait to be estimated is small. Otherwise, composite sampling will not be cost-effective.

- There are no practical difficulties that impede the selection of random samples of units.

- There are no practical difficulties in forming appropriate composites (e.g., individual samples can be adequately homogenized).

- Compositing will not affect the sample integrity.

- The likelihood of mis-classifying a composite sample is negligible. If the trait of interest is based on whether an analyte is or is not present, then this condition implies that the individual samples either should have "high" levels or (essentially) zero levels. That is, composite sampling in this context will work best when there is a clear distinction between those units with and without the trait. For instance, if a spill has occurred, then concentrations can be expected to be many times larger than the detection limits for the affected units and will be essentially zero for all unaffected units. [Garner et al. (1989) describe ways of dealing with cases where mis-classification errors may be non-negligible.]

9.3.3 What are the benefits of using composite sampling?

The benefit of using compositing for estimating a population proportion is the cost savings that can be achieved. As an example, suppose that 40 composite samples each comprised of 20 randomly selected units can be used. If the cost of an analysis is $100 and the cost of collecting and compositing samples is $5 each, then the total cost (excluding fixed, set-up costs) would be $8000 (40x$100 + 40x20x$5). If the (unknown) proportion of positive units in the population is 0.05, then the above design would be expected to produce a 95 percent confidence interval with width of ±0.021. If one chose not to use compositing, over 300 individual samples and analyses would be needed to attain about the same precision, but the cost would be substantially higher—more than $31,500 (300x$105).

9.3.4 What are the limitations of using composite sampling?

The main limitations are those implied by conditions (4) through (7) in Section 9.3.2. Of particular concern is condition (7) when analyte presence/absence is the trait of interest, since mis-classification probabilities can become non-negligible due to dilution effects. Another potential problem is that the optimal number of units to be combined into each composite, based on statistical considerations (see Section 9.3.5), may be too large to be practical [i.e., condition (5) fails to hold].

9.3.5 How do you implement composite sampling?

Let p denote the unknown proportion of the units in the population having the given trait (i.e., p is the probability of observing a positive result for a randomly selected unit in the population). To use composite sampling to estimate p requires that one choose m random samples of k units each from the population and then form m composites from the k units in each sample. Each composite is then tested for the trait. If x (which depends on k) is the number of positive test results that occur from among the m composites, then x/m can be used to estimate p^*, where p^* denotes the probability that composites of size k will test positive. Because the number of positive results, x, has a known statistical distribution (a binomial distribution with parameters m and p^*) when random samples are used, the relationship between p^* and p has been

derived. This relationship is $p=1-(1-p^*)^{1/k}$. Thus by substituting the observed x/m fraction for p^* in this equation, an estimate for p can be obtained. This estimate is satisfactory if the misclassification rates are small. Modified estimates are needed otherwise (see Garner et al., 1989).

The specific directions for implementation are given in Box 4. As indicated there, to arrive at the best combination of *m* and *k*, one needs to have some idea of the maximal p value to be expected; also, one needs to specify the precision with which p is to be estimated. Table 28 provides guidance in choosing *m* and *k* to meet the precision requirement. Note that increasing m, the number of composites, increases the precision. The table shows the optimal *k*, from a statistical precision-of-estimation standpoint, for the different *m* and *p* values. Note that this optimum *k* increases with increasing *m*. If *p* is small, then the *k* values indicated may be too large to be practical, either because of homogenization difficulties or because of concerns about dilution effects that would negate the assumption of minimal mis-classification errors. The table also gives approximate 95% confidence interval widths that can be with the optimum *k*. Composite sampling protocols for estimating p can also be derived that take into account the relative costs of sample collection, handling, compositing, and testing. The cost model is the same as in Table 33.

Table 28. Optimal k for Estimating p and Approximate Confidence Intervals for p

m	Maximum Anticipated Prevalence (p)							
	p=0.25		p=0.10		p=0.05		p=0.01	
	opt. k	~ C.I.	opt. k	~ C.I.	opt. k	~ C.I.	opt. k	~ C.I.
100	5	±0.06	14	±0.02	30	±0.012	-	
70	5	±0.07	14	±0.03	30	±0.015	-	
50	5	±0.08	12	±0.04	25	±0.018	-	
40	4	±0.09	12	±0.04	20-25	±0.021	-	
30	4	±0.11	10	±0.05	20	±0.025	-	
20	3	±0.14	8	±0.06	15	±0.03	60	±0.007
10	3	±0.22	5	±0.11	9	±0.06	35	±0.014

*Information is based on Garner et al. (1989), which contains more m & k combinations.

9.3.6 Under what conditions are other designs commonly used in conjunction with composite sampling?

Simple random samples should be used in this situation; otherwise, the relationship between p and p^* may not hold.

9.3.7 Examples

An application of composite sampling for estimating a proportion would be seen in the early stages of a large-scale investigation that encompasses a large number of sites (or batches),

Box 4. Directions for Composite Sampling for Estimating the Proportion of a Population with a Given Trait

STEP 1: Based on prior knowledge, determine an upper bound on *p*, the proportion of the population having the trait of interest. Because compositing will not be cost-effective when p exceeds 0.25, continue with Step 2 only if p is less than or equal to 0.25.

STEP 2: Determine an appropriate volume for each sample, and determine the maximum number of samples that can be composited, based on the physical nature of the samples and the trait to be assessed and on the capability to combine and homogenize the samples adequately. Denote this maximum number of samples as *K*.

STEP 3: Use the *p* value from Step 1 to identify a column in Table 28 (or an interpolated value). Within this column, consider any row in which the optimal *k* is less than *K*. Within this set of rows, select the smallest m value for which the precision is deemed acceptable. (The precision in Table 28 is expressed in terms of an approximate 95% confidence interval.)

STEP 4: Compute the cost of selecting $n = m \times k$ samples and analyzing *m* samples to determine if the composite has the trait of interest: Cost = fixed set-up costs + (cost of selecting/handling a sample) x *n* samples + (cost of forming a composite + cost of performing an analysis) x *m* analyses.

STEP 5: If the cost exceeds the available resources, consider cost-precision tradeoffs to arrive at useful *m* and *k* values. (Note that the restriction that k be less than or equal to *K* may lead to situations in which optimal combinations of *m* and *k* are not possible. The formulas in Appendix 9B can be used to calculate the precision for a given combination, and the formula in Step 4 can be used to calculate the total cost.)

STEP 6: Use a simple random sampling strategy to select *k* samples at random from the target population. (See Chapter 3, for example, for procedures for selecting a random sample.)

STEP 7: Form one composite sample as follows: physically homogenize each of the *k* samples (this may be unnecessary if each sample is of the same volume and each sample is to be included in the composite in its entirety). Obtain equal-volume aliquots of material from each of these *k* samples and physically mix the selected material and homogenize it thoroughly.

STEP 8: Repeat Steps 6 and 7 *m* times to form *m* composite samples.

STEP 9: Obtain measurements on the environmental parameters of interest for each composite sample or, if necessary, on equal-volume aliquots therefrom. Let *x* be the number of the *m* composites that have the trait and define $P=x/m$. Calculate the prevalence of the trait in the population as $\hat{p} = 1-(1-P)^{1/k}$. See Appendix 9B for estimating precision of this estimate.

where each site has a defined set of sampling locations (e.g., points defined by a fixed-size grid). The goal in such a case is to identify the subset of sites with the highest and lowest proportions (across the grid points), so that further testing or remediation actions can be taken based on the results. If the individual samples can be preserved so that subsequent aliquots can be selected, then other uses of composite sampling, such as those described in Sections 9.4 and 9.5, can be used with this strategy. A similar application would be in the context of a finite population (e.g., a field of drums with potential contamination in some), another situation in which estimating the contamination proportion via composite sampling would be the first step in an iterative study.

9.4 COMPOSITE SAMPLING AND RETESTING FOR IDENTIFYING A TRAIT

9.4.1 Overview

Another main use for composite sampling is to classify units as having or not having a trait. This use requires some of the units to be retested; hence, the identity of the units must be maintained and only aliquots (not the entire samples) of the individual sampled units composited. The particular units to be retested depend on the analytical results found for the composites. Hence the analytical results must be readily available for these composite sampling and retesting schemes to be practical. Since the goal of the composite sampling and retesting protocol in this situation is to classify each unit, as opposed to making a statistical inference about the population represented by the units, composite sampling and retesting would typically be encountered when there is a finite number of units and classification of all units is needed. This section discusses various composite sampling and retesting schemes. There are two basic cases that must be distinguished. A fundamental assumption *for both cases* is that *measurement error is very small* and does not interfere with the yes/no identification.

The first case involves a situation in which a truly binary trait (a yes/no measure) is the trait of interest. Using the schemes in this case assumes that if any unit comprising a composite has the trait, then the composite will as well. As a consequence, samples in a "negative composite" do not need additional testing, while samples in a "positive composite" do require some further testing via some scheme. In this situation, there is a clear distinction between units having and not having the trait—i.e., misclassification errors (including, for instance, impacts of dilution effects) are very small. The fundamental notion underlying this use of composite sampling is that by retesting units (either singly or in sub-composites), the identity of the positive unit(s) can be determined and if the trait is rare enough, then the composite sampling and retesting strategy will require fewer analyses than simply testing all individual units. The literature for this case is extensive, and determination of properties of various composite sampling and retesting schemes (e.g., expected cost, as a function of prevalence of the trait) is relatively straightforward, at least for the simpler schemes.

The second case involves comparing a continuous, non-negative measure x (e.g., a concentration of contaminant) with a threshold level C. (For ease of exposition, x will be referred to as a concentration.) In this case, a unit having $x \geq C$ is said to have the trait. To identify such

units when composites are employed, one obviously has to adjust the threshold level. The worst case situation would be if $x=C$ for one unit in a composite and $x=0$ for the remaining k-1 units; to identify the unit with $x=C$, one would need to compare a composite's concentration to C/k, if k samples are used to form composites. The difficulties with this case are twofold: (1) a "positive composite" does not necessarily imply that one or more of its component units is positive (i.e., there can be composites with concentrations greater than C/k that do not contain any individual units with $x \geq C$) and (2) the assumption of negligible measurement errors is likely to be untenable in many circumstances. It is much more difficult to estimate costs for a given composite sampling and retesting scheme in this case because the expected number of analyses depends on the spatial (or temporal) distribution of the underlying measure x rather than just on the prevalence of the trait.

9.4.2 Under what conditions is it appropriate to use composite sampling and retesting?

It will generally be appropriate to use composite sampling and retesting for classifying units when the following conditions hold:

1. There is a predefined set of units that are to be classified. This could be a sample from some population, but inferences are limited to the units actually within the set.

2. Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. Otherwise, composite sampling and retesting will not be cost-effective.

3. The proportion of the population having the trait of interest is small. Otherwise, composite sampling and retesting will not be cost-effective.

4. Representative aliquots from the individual units can be obtained.

5. There are no practical difficulties in forming appropriate composites from aliquots (e.g., aliquots can be adequately homogenized).

6. Compositing will not affect the sample integrity.

7. Retesting of units is feasible. In particular, the identity of the units can be maintained, and samples can be adequately preserved throughout all the potential stages of testing and retesting.

8. The likelihood of mis-classifying a composite sample is minimal. If the trait of interest is based on whether an analyte is or is not present, then this condition implies that the individual samples either should have "high" levels or (essentially) zero levels—i.e., composite sampling in this context will work best when there is a clear distinction between those units with and without the trait.

9. Analytical results are available in a timely manner. (Some retesting schemes require that test results be available sequentially; others do not. However, no composite sampling and retesting scheme will be practical if acquiring test results is a time-consuming process.)

9.4.3 What are the benefits of using composite sampling and retesting?

The benefit of composite sampling for this purpose is the cost savings that can be achieved. Substantial gains can be realized if the trait of interest is rare and the analytical costs are high relative to the costs of collecting, handling, maintaining, and compositing samples.

9.4.4 What are the limitations of using composite sampling and retesting?

The main limitations are those implied by conditions (4) through (9) in Section 9.4.2. Of particular concern is condition (8) when analyte presence/absence is the trait of interest, since mis-classification probabilities can become non-negligible due to dilution effects (which yield false negative results).

9.4.5 How do you implement composite sampling and retesting?

Box 5 gives directions for use of composite sampling and retesting to classify units when the trait of interest is a *binary trait*. Various composite sampling and retesting schemes are available. These are listed in Table 29 (see Section 9.4.7 for examples). To determine an appropriate scheme, one first needs to determine if the test procedures allow the classification results to be immediately available (see second column of table). If so, the sequential types of schemes can be considered, but if not, one of the non-sequential schemes should be used. In general, the ordering of the schemes (rows) in Table 29 is from least to most complex and from least to most cost efficient. Table 29 (column 4) also refers to Tables 30, 31, and 32, which give optimal k values and relative costs for three of the schemes. The costs are based on the assumption that the positive units occur at random. However, if available information allows grouping the more similar units (with respect to the trait) into composites, then such grouping will reduce the expected number of tests and hence the study cost.

When the trait is based on whether a *continuous, non-negative measure exceeds or does not exceed a threshold* level C, then the fundamental difficulty is that a positive composite (a composite with level above C/k) does not necessarily imply that one or more of its component units is positive (has a level above C). In this case, the choice of k must not only consider the rarity of the trait but also the detection limit (DL) and how C relates to the DL. In particular, k must be chosen to be less than C/DL, where DL is a detection limit for which the likelihood of detection is high if concentrations at that level occur. As previously noted, the determination of (relative) cost is difficult in this situation, since it is not simply a function of k and the trait's rarity but also depends on the distribution of contaminant over the site or process. Thus tables showing optimum k values cannot be given for this case.

For this second case, if test results are not immediately available, one approach is simply to test composites versus a threshold C/k and to retest all units in any positive composite. If test results are immediately available, another possible scheme for this case is the Dorfman-Sterrett Retesting scheme. This approach is similar to the curtailed exhaustive retesting scheme for binary traits (see Table 29); it relies heavily on the assumption of minimal measurement error. If Y denotes the composite-sample result and X_j, j=1,2,...,k denote potential individual-unit results, then the composite will be called positive if $kY>C$. Note that kY is expected to equal the sum of the Xs (since Y is a physically averaged mean). Thus, if $kY<C$, then all of the Xs are less than C and no further testing of the units in the composite is needed. Otherwise, testing continues sequentially until: $kY-X_1 < C$ (which implies $X_2, X_3, ..., X_k$ are all less than C), or $kY-X_1-X_2 < C$ (which implies $X_3, X_4, ..., X_k$ are all less than C), or $kY-X_1-X_2-X_3 < C$ (which implies $X_4, X_5, ..., X_k$ are all less than C), or $kY-X_1-X_2-...-X_{k-1} < C$ (which implies X_k is less than C).

Box 5. Directions for Compositing and Retesting to Classify Units Having a Rare Trait

STEP 1:	Based on prior knowledge, determine an upper bound on p, the proportion of the population having the trait of interest.
STEP 2:	Determine the maximum number of aliquots that can be composited, based on the physical nature of the samples and their volumes and any concerns about dilution effects and the capability to combine and homogenize them adequately. Denote this maximum number of aliquots as K.
STEP 3:	Decide on appropriate composite sampling and retesting scheme, based on whether classification results are available immediately (see Table 29) and whether a sequential scheme seems practical.
STEP 4:	For the given scheme, determine if the optimal k is less than K. If so, use the optimal k; if not, use $k=K$.
STEP 5:	Estimate costs for the selected k and anticipated prevalence, based on the number of units to be classified. (Formulas depend on the scheme.)
STEP 6:	If the estimated cost exceeds the available resources, then consider cost-efficiency tradeoffs to arrive at useful m and k values. (Note that the restriction that k be less than or equal to K may lead to situations in which an optimal combination of m and k is not possible.)
STEP 7:	Implement the protocol. (Note that grouping units that are more likely to have the trait into a single composite will result in fewer required analyses than the number that would occur by chance (random groupings), so that type of information, if available, should be used.) In forming each composite, thoroughly mix the material in each of the samples and obtain equal-volume aliquots from each one; then form the composite and homogenize the composite thoroughly prior to making the classification measurement.

Actually, testing of the k^{th} sample is unnecessary, so at most k tests per composite are needed. The specific algorithm for this case is given by Patil et al. (1996, p. 43).

A second approach, which also assumes that the test results are immediately available, is a Binary Split Retesting scheme. This scheme is similar to the curtailed binary split retesting scheme for binary traits; it also relies heavily on the assumption of no measurement error, since its testing-retesting strategy is based on same logic as for the prior scheme. The specific algorithm for this case is given by Patil et al. (1996, p. 44).

9.4.6 Under what conditions are other designs commonly used in conjunction with composite sampling and retesting?

Typically, when attempting to classify units, a finite population is of interest and one would use a census (e.g., all wells in a region or drums in a field). If random sampling or grid sampling is employed, then inferences are restricted to the set of selected sampling units.

Judgement can be used to form composites if there is adequate information. If the goal is to classify all units, then statistical inference from a sample to a population is not relevant. Hence, if one has knowledge about which units are more homogeneous *with respect to the likelihood of having the trait* (as opposed to simply random choices), then compositing those units believed to be most alike will be more efficient than random groupings (if the information is correct). In this situation (i.e., a situation in which stratification or ranking might be used), fewer positive composites should occur than in a simple random sample situation and hence less retesting will be required.

9.4.7 Examples

Figure 28 illustrates how several of the composite sampling and retesting schemes would perform for a case in which classification is based on a binary trait. The example involves a situation in which 32 units (numbered 1 to 32) are to be classified and for which three particular units have the trait— namely, units 6, 19, and 20. As an illustration of how to interpret Figure 28, consider the binary split algorithm. At the first stage, aliquots are used to form four composites of eight units each. The first and third composites test positive because they contain unit 6 and units 19 and 20, respectively, while the second and fourth composites test negative. Further testing is needed for the units in the first and third composites, so at stage 2 four new composites are formed, each containing four units. Since two of these composites test positive, stage 3 involves forming four new composites of two units each, two of which test positive. Finally, at stage 4, aliquots from four individual samples are tested and the three positive units are identified. The overall process involved a total of 16 tests for this example, as compared to 32 if the individual units had been tested. The figure shows that the various composite sampling and retesting schemes required from 13 to 20 tests for this example. As illustrated by this example, the usual situation would involve classification of all units in some given population—for instance,

Table 29. Identification of Composite Sampling and Retesting Schemes for Classifying Units Having a Rare Trait

Name of Scheme	Sequential test results available immediately?	Description of Procedure	Table Providing Optimal k Values	Reference for Exact Algorithm
Exhaustive Retesting (Dorfman)	No	Each composite is tested; for each composite that tests positive, all individual samples are tested (at least one is positive).	9-6	Patil et al. (1996, p. 21), based on Samuel (1978)
Curtailed (Dorfman)	Yes	Each composite is tested; for each composite that tests positive, the individual samples are tested sequentially; if k-1 have tested negative, then the k^{th} unit is taken to be positive (without testing).		Patil et al. (1996, p. 24)
Sequential Retesting (Sterrett)	Yes	Each composite is tested; for each composite that tests positive, the individual samples are tested sequentially until a positive one is found; composites of the remaining samples are formed and tested in similarly.	9-7	Patil et al. (1996, p. 27) based on Sterrett (1957).
Curtailed (Sterrett)	Yes	Like Sequential Retesting, except that if all but one of units from a positive composite have been tested and found to be negative, then testing of last unit is unnecessary (since it is known to be positive).		Patil et al. (1996, p. 29)
Binary Split Retesting	No	Each composite of *k* units is tested; for each composite that tests positive, new sub-composites containing *k*/2 units are formed and tested; this splitting and retesting is continued until all units are classified.	9-8	Patil et al. (1996, p. 32), based on Gill and Gottlieb (1974)
Curtailed Binary Split Retesting	Yes	Like the Binary Split Retesting scheme, except that only a single subcomposite of *k*/2 units (from a positive composite) is formed and tested; if it is found negative, a sub-subcomposite of *k*/4 units is formed and tested; this splitting and retesting is continued until all samples are classified.		Patil et al. (1996, p. 36), based on Gill and Gottlieb (1974)
Entropy-Based Retesting	Yes	Composites are formed sequentially; when a positive composite of k units is found, a single subcomposite containing *k*/2 units is formed; if it is positive, it is handled as in the curtailed binary split retesting scheme; but whenever a negative result occurs for the first subcomposite (or sub-subcomposite), the remaining units are grouped with those remaining to be classified and a new composite of *k* units is formed.		Patil et al. (1996, p. 39), based on Hwang (1984)

Table 30. Optimal Number of Samples per Composite for Exhaustive Retesting

Anticipated Prevalence (p)	Optimum k	Relative Cost*
0.31 and above	1	1
0.13, 0.14,..., 0.30	3	0.67, 0.70, ... , 0.99
0.07, 0.08,...,0.12	4	0.50, 0.53, ... , 0.65
0.05, 0.06	5	0.43, 0.47
0.03, 0.04	6	0.33, 0.38
0.02	8	0.27
0.01	11	0.20
0.005	15	0.14
0.001	32	0.06

* Based on random occurrence of positives. Equals $1+(1/k)+(1-p)^k$.
Source: Patil et al. (1996).

Table 31. Optimal Number of Samples per Composite for Sequential Retesting

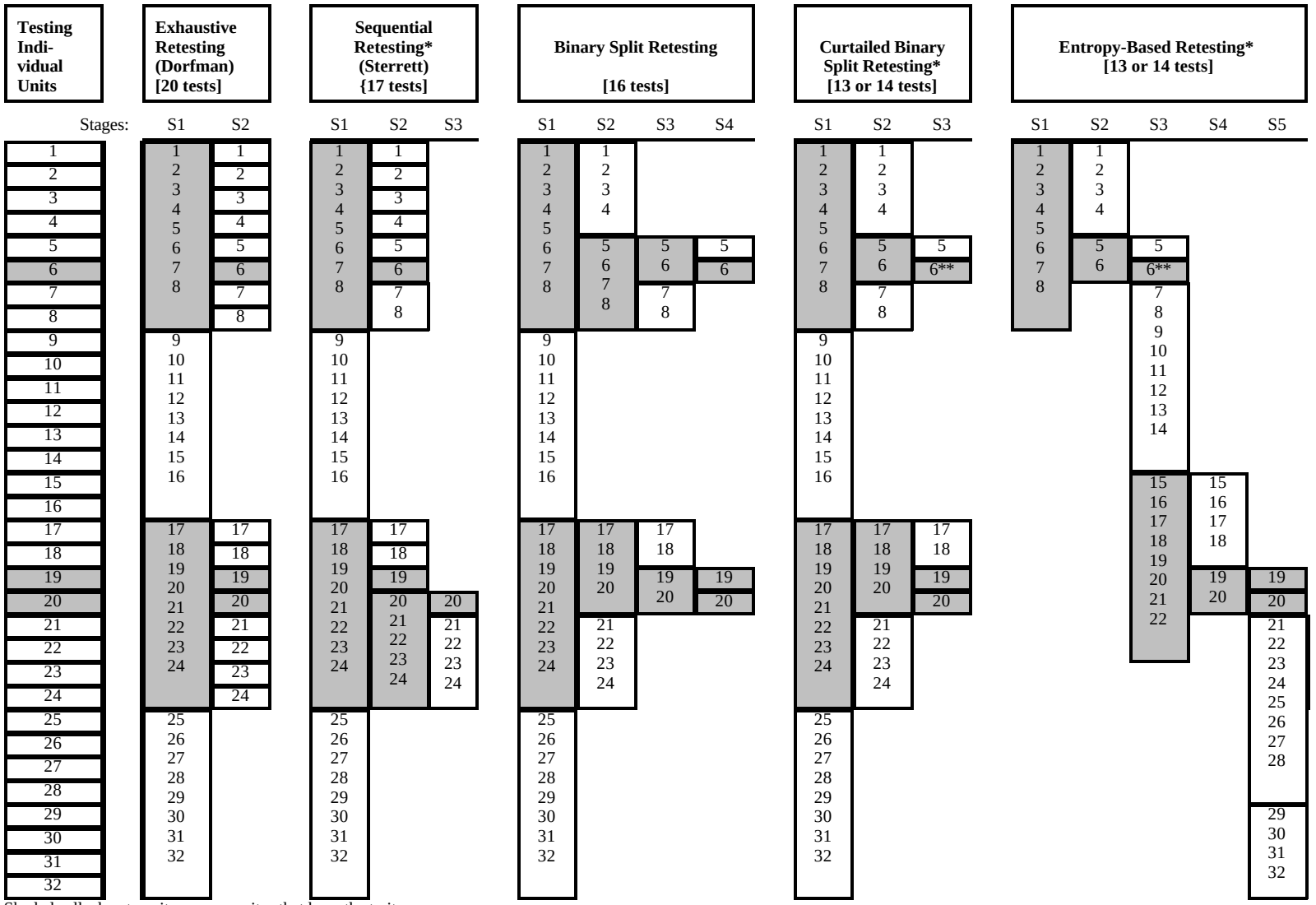
Anticipated Prevalence (p)	Optimum k	Relative Cost*
0.31 and above	1	1
0.22, 0.23,..., 0.30	3	0.843, 0.863,..., 0.992
0.13, 0.14,...,0.21	4	0.629, 0.654, ... , 0.822
0.09, 0.10, 0.11, 0.12	5	0.510, 0.541, 0.571, 0.600
0.06, 0.07, 0.08	6	0.406, 0.443, 0.478
0.05	7	0.367
0.04	8	0.324
0.03	9	0.276
0.02	11	0.221
0.01	15	0.152
0.005	21	0.106
0.001	45	0.046

* Based on random occurrence of positives. Equals $2 - q + (3q-q^2)/k - [1-q^{k+1}]/(kp)$ where $q=1-p$. Source: Patil et al. (1996).

Table 32. Optimal Values of k (k=1 to 16) for Given m (m=2,4,6,...16) and p for Binary Split Retesting

Anticipated Prevalence (p)	Value of m (number of initial composites of size k)							
	2	**4**	**6**	**8**	**10**	**12**	**14**	**16**
>0.38	1	1	1	1	1	1	1	1
0.28 - 0.38	2	2	2	2	2	2	2	2
0.27	2	2	2,3	2	2	2	2	2
0.26	2	2	3	2	2	2	2	2
0.25	2	2	3	2,3	2,3	2,3	2,3	2
0.24	2	2,4	3	3	3	3	3	2,3
0.20 - 0.23	2	4	3	3	3	3	3	3
0.19	2	4	3	3,4	3,4	3,4	3	3,4
0.18	2	4	3	4	4	4	3,4	4
0.17	2	4	3,4	4	4,5	4	4	4
0.16	2	4	4,6	4	5	4	4,5	4
0.15	2	4	6	4	5	4	5	4
0.14	2	4	6	4	5	4,5	5	4
0.13	2	4	6	4	5	5,7	5	4,7
0.12	2	4	6	4,8	5	7	5,7	7
0.11	2	4	6	8	5	7	7	7,8
0.10	2	4	6	8	5,10	7	7	8,11
0.09	2	4	6	8	10	7	7	11
0.08	2	4	6	8	10	7,12	7	11,16
0.00 - 0.07	2	4	6	8	10	12	14	16

Source: Patil et al. (1996).



Shaded cells denote units or composites that have the trait.

* This scheme assumes tests are conducted in sequence and that test results are available immediately. ** Test does not actually have to be conducted.

Figure 28. Illustration of Retesting Schemes for Classifying Units When 3 of 32 Units are Positive

classification of all drinking water wells in a given region to determine those that have animal waste contamination.

An application of the binary split algorithm for testing drinking water wells for pesticide contamination is described by Natarajan and Rajagopal (1993). Spiked samples were also used to assess the performance of the procedure.

9.5 COMPOSITE SAMPLING AND RETESTING FOR IDENTIFYING EXTREME SAMPLING UNITS

9.5.1 Overview

All units in a given population might be exhaustively tested if the interest is in identifying the maximum value (e.g., concentration of contaminant *X*) that occurs among those units (e.g., *n* drums or *n* spatial grid points). However, under certain circumstances, composite sampling coupled with the retesting of some units in one or more of the composites with the highest levels can reveal the unit with the maximum level—and at a substantially reduced cost. Similar composite sampling and retesting strategies can be used to identify the second highest unit, the third highest unit, etc.—i.e., to determine the upper percentiles of the population. This use of compositing has recently received some theoretical attention; real examples, however, do not appear to exist.

9.5.2 Under what conditions is composite sampling and retesting appropriate to use to identify extreme sampling units ?

It will generally be appropriate to use composite sampling and retesting for determining the unit with maximum value (and the value itself) when the following conditions hold:

1. There is a predefined set of units that are to be evaluated.
2. Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. (Otherwise, composite sampling and retesting will not be cost-effective.)
3. Representative aliquots from the individual units can be obtained.
4. There are no practical difficulties in forming appropriate composites from aliquots (e.g., aliquots can be adequately homogenized).
5. Compositing will not affect the sample integrity.
6. Retesting of units is feasible. In particular, the identity of the units can be maintained, and samples can be adequately preserved throughout all the potential stages of testing and retesting.
7. Measurement error, relative to the range of *X* concentrations, must be minimal. This assumption is necessary because the tests on individual units or on composites of a given size have to be properly ordered.
8. Analytical results are available in a timely manner. (Otherwise, the retesting schemes will be impractical.)

9.5.3 What are the benefits of using composite sampling and retesting to identify extreme sampling units?

Costs can be substantially less than using exhaustive testing.

9.5.4 What are the limitations of using composite sampling and retesting to identify extreme sampling units?

The main limitations are those implied by conditions (3) through (8) in Section 9.4.4, especially condition (7), the assumption of negligible measurement error.

9.5.5 How do you implement composite sampling and retesting to identify extreme sampling units?

Several methods have been proposed for using composite sampling and retesting to identify a maximum. These include the following:

- Casey et al. (1985) proposed a simple method for predicting a maximum in which searching for the maximum is restricted to the composites having the highest levels. The simplest strategy would involve retesting only those units in the composite that had the highest level; this approach obviously cannot guarantee that one finds the unit with the maximum level.

- Gore and Patil (1994) proposed a *sweep-out method for identifying a maximum*. This method is based on the idea that if the composite with the highest observed value has a level that is "sufficiently" bigger than the level for the next highest composite, then one can be assured (apart from possible measurement error concerns) that the unit with highest level is in the composite with the highest level; hence retesting of the individual samples in that composite will reveal the unit with the maximum (with certainty, if there is no measurement error). However, if the "gap" in observed levels for these two composites is not "sufficiently" large, then retesting of units in some other composites may be needed to identify the one with the maximum level. The algorithm relies on the fact the maximum individual value within a composite consisting of k units must fall in the interval $[Y, kY]$, where Y is the value observed for the composite. This property assumes no measurement error is involved; it is true because the maximum must be larger than the average, Y, and smaller than the total, kY. Thus if the units in the composite with the highest Y are each individually tested and the maximum value is found to be Z, then the composite measurements for each of the other composites can be compared with Z to determine if they might contain the maximal unit: only composites where $kY > Z$ can contain the maximal unit and hence only the units within those composites require further testing. The explicit algorithm for this method is given by Patil et al. (1996, p. 51).

- An extension of Method B is a *sweep-out method for identifying upper percentiles*, which is described by Patil et al. (1996, p. 53).

- Two other methods, which are extensions of Method B and which are described by Gore et al. (1996), are a *locally-sequential sweep-out method* and *globally-sequential sweep-out method* for identifying a maximum.

9.5.6 Under what conditions are other designs commonly used in conjunction with composite sampling and retesting to identify extreme sampling units?

Most commonly, there is a finite population of interest. A grid sample might be viewed as a surrogate "census" if interest is confined to the "sample." With a maximum being the parameter of interest, no statistical inference beyond the sample (whatever it may be) is generally warranted.

9.5.7 Examples

No real environmental applications appear to exist for this use of composite sampling. However, under the conditions described in Section 9.5.2, composite sampling would appear to have great cost savings potential.

APPENDIX 9-A

COST AND VARIANCE MODELS FOR COMPOSITE SAMPLING AIMED AT ESTIMATING A MEAN

Table 33 provides information on the costs and variances encountered when estimating a mean either when a simple random sample is used alone or when composite sampling is coupled with random sampling. The first few rows of Table 33 list the typical cost components that occur. In the first (simple random sampling) case, it is assumed that n samples are individually analyzed, while in the second case, it is assumed that $n = km$ samples are selected, where m is the number of composites (each comprised of k samples) analyzed. As indicated in the table, there are three cost components considered, in addition to the fixed set-up cost, and the total cost is determined by adding these components together, appropriately weighted by the number of samples or analyses.

The lower portion of Table 33 defines the pertinent variance components and indicates how they combine to yield the variance of an estimated mean for the simple random sampling and the simple random sampling with compositing cases. For the latter case, the relevant components are (1) the inherent variability of the site or process, (2) the variability associated with collecting and handling the samples, (3) the variability associated with the compositing process, and (4) the variability associated with all aspects of the measurement process. In most circumstances, (1) and (2) cannot be separately estimated and are therefore treated as one.

From a statistical standpoint, an optimal k can be determined as follows, if the cost of compositing (C_C) can be ignored and the additional variance due to compositing (V_C) can be considered negligible:

$$opt.\ k = \sqrt{\frac{V_I\ C_M}{V_M\ C_S}}$$

The optimum is more difficult to determine if either the cost of compositing or the additional variance due to compositing is not negligible, since these components depend on k. However, the formulas in the table permit one to calculate the costs and variances for any combination of k and n.

If an optimal k value is too large for practical implementation, the maximum practical k would typically be employed. Further, a smaller k value might be employed if a sufficiently large m is needed to estimate adequately the precision of the estimated mean. With the "optimal" k, one would expect to obtain a more precise estimate but may not be able to estimate the precision well. Hence in those cases where a good estimate of the precision is also needed—for example, if confidence interval estimates are to be constructed or hypothesis tests are to be performed (since these rely on both an estimate of the mean and its standard error)—one should consider using a smaller k if that is the only way to achieve an adequate number of composites, m, for estimating the standard error.

Box 1 of Section 9-2 provides another example of how the information in Table 33 can be used. It compares the cost and precision for a composite sampling approach versus an simple random sampling, under the assumptions that C_C and V_C are negligible. The cost comparison is derived by dividing the total cost in the last column by the total cost in the first column. After some algebra, this relative cost ratio can be shown to be

$$RC = \frac{C_0 + nC_S\,[1 + \frac{1}{k}\frac{C_M}{C_S}]}{C_0 + nC_S\,[1 + \frac{C_M}{C_S}]}$$

If the fixed-cost component is ignored (as in Box 1), then this equation shows that the relative cost, RC, reduces to a function of k and the C_M/C_S ratio (i.e., only the part inside the brackets remains). The precision comparison given in Box 1 was derived from the information in the last row of the table; after some algebra, and the assumption that $f_C = 0$, the relative variance (RV) of composite sampling versus an simple random sampling without compositing is

Table 33. Components of Cost and Variance for Random Samples - With and Without Composite Sampling

Component	Random (or Grid) Sample without Composite Sampling	Random (or Grid) Sample with Composite Sampling
COSTS		
Fixed cost	C_0	C_0
Per-unit sampling cost	C_S (x n samples)	C_S (x n samples)
Per-composite compositing cost	0	C_C (x m analyses)
Per-unit measurement cost	C_M (x n analyses)	C_M (x m analyses)
Total Cost	$C_0 + nC_S + nC_M$	$C_0 + nC_S + mC_{C+} mC_M$
VARIANCES OF CONTAMINANT X		
Inherent variability*	V_I	V_I
Per-composite compositing variability	0	V_C $(= f_C V_I)$**
Per-unit measurement variability	V_M	V_M
Total variance of an observed measurement	$V_I + V_M$	$[V_I (1+f_C)]/k + V_M$
Variance of mean of all measurements	$[V_I + V_M]/n$	$\{[V_I (1+f_C)]/k + V_M\}/m =$ $[V_I (1+f_C)]/n + V_M/m =$ $[V_I (1+f_C)] + kV_M]/n$

n = number of samples
m= number of measurements (= number of composite samples)
k = number of samples/composite
* Includes variability associated with collecting and handling of samples.
** This component is generally small. It is considered to be proportional to the V_I for the following reason: if the individual samples comprising a composite all have essentially the same X concentration (i.e., a small V_I) , then failure to have exactly equal volumes from each individual sample or failure to homogenize them adequately will result in relatively little error in the X level of the composite; on the other hand, these sorts of procedural flaws will lead to larger errors if the individual samples exhibit large variation (i.e., a small V_I).

APPENDIX 9-B

COMPOSITE SAMPLING FOR ESTIMATING A POPULATION PROPORTION

Details for using composite sampling to estimate a proportion are provided by Garner et al. (1989). It is assumed that there is a large (essentially infinite) population of units and that n are selected at random. Suppose $m=n/k$ composites of k units each are formed (by random groupings). Let the probability of a composite testing positive be denoted by p^* and let the probability of an individual unit testing positive be denoted by p (p is the proportion of the population having the trait and is the quantity that is to be estimated). Since the probability of a composite testing negative is the same as the probability that all k units will test negative, then

$$1-p^* = (1-p)^k \quad \text{or} \quad p^* = 1 - (1-p)^k \qquad \text{(Eq. 9B-1)}$$

Solving for p yields

$$p = 1 - (1-p^*)^{1/k} \qquad \text{(Eq. 9B-2)}$$

This equation offers a simple way to estimate p: If x of the m composites are found to have the trait, then substituting x/m for p^* in this equation will yield an estimate of p. This estimate is a maximum likelihood estimate; it is not an unbiased estimate, however.

Since the number of positive composites has a binomial distribution with parameters m and p^*, one can use tables for binomial confidence limits to derive corresponding confidence limits for p. [Binomial confidence limits are available in many standard statistical texts; Gilbert, 1987, also provides some tables and formulas (pp. 142-143.)] For example, suppose composites comprised of 10 units each were used. If 10 out of 30 of these composites tested positive, then the point estimate for p^* is $x/m=0.3333$ and a 95 percent confidence interval for p^* is [0.1729, 0.5280]. Using x/m and these interval end points in Eq. (8B-2) yields a point estimate of p equal to 0.0397 and a corresponding interval estimate of [0.0188, 0.0723].

If the mis-classification rates are not negligible, and one knows them, then one can substitute

$$\frac{\frac{x}{m} - \alpha}{1 - \alpha - \beta}$$

for the p^* appearing in Eq. (8B-2), where α and β are the false positive and false negative rates, respectively.

GLOSSARY OF TERMS

biased sampling - see "judgmental sampling."

coefficient of variation (CV) - a unit-free measure of variability. It is the ratio of the standard deviation divided by the mean and is often expressed as a percentage.[4] As such, the CV can be thought of as the standard deviation expressed as a percentage of the mean. Hence, the sample CV is given by

$$cv = \frac{s}{\bar{x}}$$

and the population CV is given by

$$CV = \frac{\sigma}{\mu}$$

confidence interval - an interval estimate of a population parameter with a known probability that the population value of interest will be included in the interval. For example, a 100 (1 - α)% confidence interval estimate of a population parameter, Θ, (e.g., a population mean or proportion) is an interval that will contain the true value of Θ in 100 (1 - α)% of all samples that could be selected with a given sampling design.

directed sampling - see "judgmental sampling."

double sampling - measuring a characteristic on a sample that is positively correlated with the outcome of interest, with a subsample of those units selected for measuring the primary outcome. The relationship between the two variables is used to generate the data needed for the primary outcome variable.

hypothesis test - a statistical procedure for determining if a sample provides sufficient evidence to reject one statement regarding the population of interest (the null hypothesis) in favor of an alternative statement (the alternative hypothesis). The null hypothesis is considered the "baseline" condition and will be rejected in favor of the alternative hypothesis only when there is overwhelming evidence the null cannot be true.

judgmental sampling - use of professional judgment to choose sampling locations.

mean (sample) - the mean of a sample of n observations—$x_1, x_2, ..., x_n$—is simply the average of those n observations and is defined as $\bar{x}$.

[4]Because the coefficient of variation is a percentage of the mean, it is not a useful measure of variability when the mean is zero or close to zero.

mean (population) - the mean value of a variable, x, over a finite population of N units—x_1, x_2, ..., x_N—is simply the average of those N population values and is defined as μ.

population parameter - a characteristic based on or defined in terms of all units in a population.

probability-based sample - a sample selected in such a manner that the probability of being included in the sample is known for every unit on the sampling frame.

professional judgment - knowledge, experience or other expertise that allows a professional to come to conclusions about an environmental data collection activity that will improve the quality of the data.

sample - a set of units or elements selected from a larger population, typically to be observed for making inferences regarding that population.

sampled population - the set of units or elements from which a sample was selected, i.e., the units that had a chance of being included in the sample.

sample size - the number of sample units to be collected.

sample support - the portion of the sampling unit that is being extracted in the field and that is subjected to the measurement protocol—the area or volume that a single sample is supposed to represent.

sampling design - a description of the sample collection plan that specifies the number, type, and location (spatial or temporal) of sampling units to be selected for measurement.

sampling frame - the list from which a sample of units or elements is selected.

sampling unit - the members of a population that may be selected for sampling.

standard deviation - the square root of the variance.

standard error - the standard error of a sample statistic, θ, (say a sample mean or proportion) is the standard deviation of the values of that statistic over repeated samples using the same sampling design (e.g., stratified simple random sampling) and the same sample size, n.

statistic - a quantity calculated from the values in a sample (e.g., the sample mean or sample variance)**.**

target population - the set of all units or elements (e.g., barrels of waste or points in time and/or space) about which a sample is intended to produce inferences.

variance (sample) - the variance of a sample of *n* observations—x_1, x_2, ..., x_n—is simply the average of the squared differences between the individual observations and the mean, divided by (n - 1).

variance (population) - the variance of a finite population of *N* values—x_1, x_2, ..., x_N—is simply the average of the squared differences between the individual observations and the mean.

REFERENCES

ANSI/ASQC E4-1994. *Specifications and Guidelines for Environmental Data Collection and Environmental Technology Programs*. American National Standards Institute (ANSI) and American Society for Quality Control (now American Society for Quality).

Barabesi, L. 1998. The computation of the distribution of the sign test statistic for ranked-set sampling, Communications in Statistics-Simulation 27(3):833-842.

Barcelona, M. J. 1988. Overview of the sampling process. Chapter 1 in *Principles of Environmental Sampling*, L.H. Keith, ed. American Chemical Society, Washington, DC.

Bohn, L.L. and D.A. Wolfe. 1992. Nonparametric two-sample procedures for ranked-set samples data, Journal of the American Statistical Association 87:552-561.

Bohn, L.L. and D.A. Wolfe. 1994. The effect of imperfect judgment ranking on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic, Journal of the American Statistical Association 89:168-176.

Cailas, M. D., R.G. Kerzee, E.K. Mensah, K.G. Croke, and R. R. Swager. 1995. A proposed methodology for an accurate estimation of the total amount of materials recycled. Resources, *Conservation and Recycling* 15, 123-131.

Casey, D. B., P.N. Nemetz, and D. Uyeno. 1985. Efficient search procedures for extreme pollutant values. *Environmental Monitoring and Assessment* 5:165-176.

Christman, M.C. 1997. Efficiency of adaptive sampling designs for spatially clustered populations. *Environmetrics* 8:145-166.

Christman, M.C. 2000. A review of quadrant-based sampling of rare, geographically clustered populations. *Journal of Agricultural, Biological, and Environmental Statistics* 5(2):168-201.

Cressie, N.A. 1993. *Statistics for Spatial Data*. John Wiley & Sons, New York.

Cochran, W.G., and G.M. Cox. 1957. *Experimental Designs*. John Wiley & Sons, New York.

Cochran, W.G. 1963. *Sampling Techniques*, 2nd ed. John Wiley & Sons, New York.

Cochran, W.G. 1977. *Sampling Techniques*, 3rd ed. John Wiley & Sons, New York.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillside, NJ.

Conover, W.J. (1980). *Practical Nonparametric Statistics*. Wiley: New York

Cox, D.D., Cox, L.H, and K.B. Ensor. 1995. Spatial Sampling and the Environment. *National Institute of Statistical Sciences Technical Report #38*, Research Triangle Park, NC.

Davidson, J.R. 1995. *ELIPGRID - PC's User's Manual*. ORNL/TM-13103, Oak Ridge National Laboratory-Grand Junction, Grand Junction, CO. Software available for download at: http://es.epa.gov/quality

Desu, M.M. and D. Raghavarao. 1990. *Sample Size Methodology*. Academic Press, Boston.

Dorfman, R. 1943. The detection of defective members of large populations. *Annals of Mathematical Statistics* 14:436-440.

Elder, R.S., W.O. Thomson, and R.H. Myers. 1980. Properties of composite sampling procedures. *Technometrics* 22(2):179-186.

Entz, T. and C. Chang. 1991. Evaluation of soil sampling schemes for geostatis analyses: A case study for the soil bulk density. *Canadian Journal of Soil Science* 17:165-176.

Garner, F.C., M.A. Stapanian, and L.R. Williams. 1988. Composite sampling for environmental monitoring. Chapter 25 in *Principles of Environmental Sampling*, L.H. Keith, ed. American Chemical Society, Washington DC, pp. 363-374.

Garner, F.C., M.A. Stapanian, E.A. Yfantis, and L.R. Williams. 1989. Probability estimation with sample compositing techniques. *Journal of Official Statistics* 5:365-374.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.

Gilbert, R.O. 1995. Ranked Set Sampling. *DQO Statistics Bulletin, Statistical Methods for the Data Quality Objectives Process*, Volume 1, Number 1. PNL-SA-26377. Pacific Northwest National Laboratory, Richland, WA.

Gill, A. and Gottlieb. 1974. The identification of a set by successive intersections. *Information and Control* 24:20-35.

Gore, S.D. and G.P. Patil. 1994. Identifying extremely large values using composite sample data. *Environmental and Ecological Statistics* 1:227-245.

Gore, S.D., G.P. Patil, and C. Taillie. 1996. Identification of the largest individual sample value using composite sample data and certain modifications to the sweep-out method. *Environmental and Ecological Statistics* 3:219-234.

Guenther, W.C. 1981. Sample size formulas for normal theory *T* tests." *The American Statistician* 35(4):243-244.

Hettmansperger, T.P. 1995. The ranked-set sample sign test, *Nonparametric Statistics* 4:263-270.

Hollander, M. and D.A. Wolfe. 1999. *Nonparametric Statistical Methods*. John Wiley and Sons, New York.

Hwang, F.K. 1984. Robust group testing. *Journal of Quality Technology* 16:189-195.

Johnson, G.D., B.D. Nussbaum, G.P. Patil and N.P. Ross. 1996. Designing cost-effective environmental sampling using concomitant information, *Chance* 9(1):4-11.

Kaul, A., G.P. Patil and C. Taillie. 1995. *Unequal Allocation Models for Ranked Set Sampling with Skew Distributions*, Technical Report Number 94-0930, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA

Koti, K.M. and G.J. Babu. 1996. Sign test for ranked-set sampling, Communications in Statistics-Theory and Methods 25(7):1617-1630.

Kotz, S., Johnson, N.L., and Read, C.B. (1982). *Encyclopedia of Statistical Sciences*. Wiley: New York, pp. 234-235.

Kraft, K.M, D.H. Johnson, J. M. Samuelson, and S.H. Allen. 1995. Using Known Populations of Pronghorn to Evaluate Sampling Plans and Estimators. *Journal of Wildlife Management* 59(1), 129-137.Lancaster, V. and S. Keller-McNulty. 1998. Composite sampling, part I. *Environmental Testing and Analysis* 7(4):15ff.

Li, F. and J. Chaplin. 1998. Evaluation of large field sampling methods for crop residue coverage measurement. *Transactions of the American Society of Agricultural Engineers* 41(3):645-651.

Mace, A.E. 1964. *Sample Size Determination*. Reinhold Publishing Company, New York.

Midwest Plan Service. 1992. *Conservation Tillage Systems and Management*. MWPS-45. MidWest Plan Service. Iowa State University, Ames, IA.

McIntyre, G.A. 1952. A method for unbiased selective sampling, using ranked sets, Australian Journal of Agricultural Research 3:385-390.

Mode, N.A., L. C. Conquest, and D. A. Marker. 1999. Ranked set sampling for ecological research: accounting for the total costs of sampling. *Environmetrics* 10:179-194.

Muttlak, H.A. 1995. Parameters estimation in a simple linear regression using rank set sampling, *Biometrical Journal* 37:799-810.

Muttlak, H.A. 1996. Estimation of parameters for one-way layout with rank set sampling, *Biometrical Journal* 38:507-515.

Myers, J.C. 1997. *Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping*. Van Nostrand Reinhold, New York.

Natarajan, U. and R. Rajagopal. 1993. Sample compositing. *Environmental Testing and Analysis* 2(3):54ff.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97:558-606.

Noether, G.E. 1987. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* 82(398):645-647.

Nussbaum, B.D. and B.K. Sinha. 1997. Cost effective gasoline sampling using ranked set sampling. *Proceedings of the Section on Statistics and the Environment*. American Statistical Association, pp. 83-87. American Statistical Association, Alexandria, VA.

Patil, G.P., A.K. Sinha and C. Taillie. 1994. Ranked set sampling, Handbook of Statistics 12, Environmental Statistics, pp. 167-200, (G.P. Patil and C.R. Rao, editors), North-Holland, New York, NY

Patil, G.P., S.D. Gore, and G.D. Johnson. 1996. *EPA Observational Economy Series, Volume 3: Manual on Statistical Design and Analysis with Composite Samples* (Draft). Technical Report No. 96-0501, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University.

Perez, A., and J.J. Lefante. 1997. Sample size determination and the effect of censoring when estimating the arithmetic mean of a lognormal distribution. *Communications in Statistics, Theory and Methods* 26 (11):2779-2801.

Pitard, F.F.. 1993. *Pierre Gy's Sampling theory and Sampling Practice*, 2nd ed. CRC Press, Boca Raton, FL.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 1992. *Numerical Recipes in Fortran 77: The Art of Scientific Computing.* Cambridge University Press, Cambridge, MA.

Satterthwaite, F.E. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2:110-114.

Seber, G.A.F. and S.K. Thompson. 1994. Environmental Adaptive Sampling. *Handbook of Statistics, Vol. 12 (Environmental Sampling)*. G.P. Patil and C.R. Rao, Eds. Elsevier Science B.V., New York.

Starks, T.H. 1986. Determination of support in soil sampling. *Mathematical Geology* 18(6):529-537.

Stehman, S.V. and Overton, S. (1994), Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling, *Journal of the American Statistical Association* 89, 30-43.

Stokes, S.L. and T.W. Sager. 1988. Characterization of a ranked-set sample with application to estimating distribution functions, Journal of the American Statistical Association 83:374-381.

Samawi, H.M. and H.A. Muttlak. 1996. Estimation of ratio using rank set sampling, *Biometrical Journal* 38:(6):753-764.

Sterrett, A. 1957. On the detection of defective members of large populations. *Annals of Mathematical Statistics*, 28: 1033-1036.

Thompson, S. K. 1990. Adaptive Cluster Sampling. *Journal of the American Statistical Association* 85:412.

Thompson, S. K., 1992. *Sampling*. John Wiley & Sons, New York.

Thompson, S.K. and G.A.F. Seber. 1996. *Adaptive Sampling*. John Wiley & Sons, New York.

U.S. Environmental Protection Agency. 1986. *Test Methods for Evaluating Solid Waste: Physical/Chemical Methods 3rd ed*. 4 vols. 530/SW-846. November. Office of Policy, Planning and Evaluation, Washington, DC.

U. S. Environmental Protection Agency. 1989. Methods for Evaluating the Attainment of Cleanup Standards, Vol 1. Soils and Solid Media. PB89-234959. Washington, DC.

U.S. Environmental Protection Agency. 1992. *Statistical Methods for Evaluating the Attainment of Cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media*. NTIS # PB-94-176-831. Office of Policy Planning and Evaluation, Washington, DC.

U.S. Environmental Protection Agency. 1994. *Guidance for the Data Quality Objectives Process (QA/G-4).* EPA/600/R-96/055. September. Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency, 1996a. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis (QA/G-9).* EPA/600/R-96/084. Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency. 1996b. *Soil Screening Guidance: User's Guide.* Publication 9355.4-23. Office of Solid Waste and Emergency Response, Washington, DC.

U.S. Environmental Protection Agency. 1998a. *Quality Assurance Guidance for Conducting Brownfields Assessments*. EPA/540/R-98/038. Office of Solid Waste and Emergency Response, Washington, DC.

U.S. Environmental Protection Agency. 1998b. *Guidance for Quality Assurance Project Plans (QA/G-5).* EPA/600/R-98/018. Washington, DC.

Wang, X.J. and F. Qi. 1998. The effects of sampling design on spatial structure analysis of contaminated soil. *The Science of the Total Environment* 224, 29-41.

Wolter, K.M. 1984. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* 79, 781-790.

Yfantis, A. A., G. T. Flatman, and J.V. Behar. 1987. Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology* 19, 183-205.